# *Shadows of the Mind*

## *A Search for the Missing Science of Consciousness*

ROGER PENROSE

*Rouse Ball Professor of Mathematics*
*University of Oxford*

OXFORD UNIVERSITY PRESS
*New York Oxford*

РОДЖЕР ПЕНРОУЗ

# ТЕНИ РАЗУМА

## *В ПОИСКАХ НАУКИ О СОЗНАНИИ*

Перевод с английского
А. Р. Логунова и Н. А. Зубченко


Москва ♦ Ижевск
2005

УДК 519.7, 530.1



**Пенроуз Р.**
Тени разума: в поисках науки о сознании. — Москва-Ижевск: Институт компьютерных исследований, 2005. — 688 с.

Книга знаменитого физика о современных подходах к изучению деятельности мозга, мыслительных процессов и пр. Излагаются основы математического аппарата — от классической теории (теорема Гёделя) до последних достижений, связанных с квантовыми вычислениями. Книга состоит из двух частей: в первой части обсуждается тезис о невычислимости сознания, во второй части рассматриваются вопросы физики и биологии, необходимые для понимания функционирования реального мозга.

Для широкого круга читателей, интересующихся наукой.

**ISBN 0-19-510646-6 (англ.)**
**ISBN 5-93972-457-4 (рус.)**



**http://rcd.ru**
**http://ics.org.ru**

# ОГЛАВЛЕНИЕ

# ПРЕДИСЛОВИЕ

Эту книгу можно считать, в некотором смысле, продолжением «*Нового разума короля*»[1] (далее — НРК). То есть я и в самом деле намерен продолжить развитие темы, начатой в НРК, однако излагаемый здесь материал можно рассматривать и совершенно независимо от предыдущей книги. Отчасти необходимость в повторном обращении к предмету первоначально возникла из желания дать как можно более обстоятельные ответы на множество вопросов и критических замечаний, которыми самые разные люди отреагировали на рассуждения и доказательства, представленные в НРК. Тем не менее, тема новой книги представляет собой совершенно самостоятельное исследование, а предлагаемые здесь идеи отнюдь не ограничиваются рамками, установленными в НРК. Одну из главных тем НРК составило мое убеждение в том, что, используя сознание, мы способны выполнять действия, не имеющие ничего общего с какими бы то ни было вычислительными процессами. Однако в НРК эта идея была представлена лишь как осторожная гипотеза; имелась также некоторая неопределенность относительно того, какие именно типы процедур следует включать в категорию «вычислительных процессов». На страницах же этой книги, как мне представляется, читатель найдет гораздо более последовательное и строгое обоснование приведенного выше общего утверждения, причем представляемое обоснование оказывается применимо ко всем типам вычислительных процессов, какие только можно вообразить. Кроме того, здесь имеется и существенно более правдоподобное (нежели это было возможно во времена НРК) предположение относительно механизма церебральной активности, посредством которого наше управляемое сознанием поведение может основываться на какой-либо физической активности невычислительного характера.

[1] *The Emperor's New Mind.* (Не так давно книга была переведена на русский язык: Пенроуз Р. *Новый ум короля*, М.: Едиториал УРСС, 2003.) — *Прим. перев.*



Упомянутое обоснование проводится по двум различным направлениям. Одно из них по сути своей негативно; здесь я решительно выступаю против широко распространенного мнения, согласно которому нашу сознательную мыслительную деятельность — во всех ее разнообразных проявлениях — можно, в принципе, адекватно описать в рамках тех или иных вычислительных моделей. Другое направление моих рассуждений можно счесть позитивным — в том смысле, что оно предполагает подлинный поиск (разумеется, в рамках необходимости придерживаться строгих и неопровержимых научных фактов) инструментов, позволяющих описываемому в научных терминах мозгу применять для осуществления требуемой невычислительной деятельности тонкие и по большей части нам пока не известные физические принципы.

В соответствии с этой дихотомией, представленная в книге аргументация разбита на две части. В первой части содержится всестороннее и обстоятельное исследование, результаты которого самым решительным образом подтверждают мой тезис о том, что сознание, в его конкретном проявлении человеческого «понимания», делает нечто такое, чего простые вычисления воспроизвести не в состоянии. Причем под термином «вычисления» здесь подразумеваются как процессы, реализуемые системами «нисходящего» типа, действующими в соответствии с конкретными и прозрачными алгоритмическими процедурами, так и процессы, реализуемые системами «восходящего» типа, которые программируются не столь жестко и способны вследствие этого к обучению на основании приобретенного опыта. Центральное место в рассуждениях первой части занимает знаменитая теорема Гёделя; приводится также подробнейшее рассмотрение следствий из этой теоремы, имеющих отношение к нашему случаю. Подобное изложение существенно расширяет аргументацию, представленную сначала самим Гёделем, а позднее Нагелем, Ньюменом и Лукасом; кроме того, здесь же я постарался по возможности обстоятельно ответить на все известные мне возражения. В этой связи приводятся также подробные доказательства невозможности достижения системами восходящего (равно как и нисходящего) типа подлинной разумности. В заключение делается вывод о том, что сознательное мышление и в самом деле должно включать в

себя процессы, которые с помощью одних лишь вычислительных методов невозможно даже адекватно *смоделировать*; еще менее способны вычисления, взятые сами по себе, обусловить какое бы то ни было сознательное ощущение или желание. Иными словами, разум, по всей видимости, представляет собой такую сущность, которую никоим образом невозможно описать посредством каких бы то ни было вычислений.

Во второй части мы обратимся к физике и биологии. Хотя отдельные звенья цепи наших умозаключений и носят здесь явно более предположительный характер, нежели строгие доказательства первой части, мы все же попытаемся разобраться, каким именно образом в пределах действия научно постижимых физических законов может возникать подобная невычислимая активность. Необходимые фундаментальные принципы квантовой механики излагаются начиная с самых азов, так что от читателя не требуется какого бы то ни было предварительного знакомства с квантовой теорией. Приводится достаточно глубокий анализ некоторых загадок и парадоксов квантовой теории с привлечением целого ряда новых примеров, графически иллюстрирующих роль нелокальности и контрфактуальности, а также некоторых весьма сложных проблем, связанных с квантовой сцепленностью. Я глубоко убежден — и готов свою убежденность обосновать — в необходимости фундаментального пересмотра (на определенном, четко обозначенном уровне) наших сегодняшних квантовомеханических воззрений. (Высказываемые здесь соображения весьма близки к идеям, недавно опубликованным Гирарди, Диози и др.) Следует отметить, что со времен НРК в этом отношении произошли существенные изменения.

Я полагаю, что именно на этом уровне в действие должна вступать физическая невычислимость — условие, необходимое для объяснения невычислимости деятельности сознания. В соответствии с этим предположением я должен потребовать, чтобы уровень, на котором становится значимой упомянутая физическая невычислимость, играл особую роль и в функционировании мозга. Именно в этом пункте мои нынешние предположения наиболее существенно расходятся с теми, что были высказаны в НРК. Я утверждаю, что, хотя сигналы нейронов и могут вести себя как детерминированные в классическом смысле события, управление синаптическими связями между нейронами происходит на более глубоком уровне, т. е. там, где можно ожидать



наличия существенной физической активности на границе между квантовыми и классическими процессами. Выдвигаемые мною специфические предположения требуют возникновения внутри микроканальцев цитоскелета нейронов макроскопического квантовокогерентного поведения (в точном соответствии с предположениями Фрёлиха). Иначе говоря, я полагаю, что упомянутая квантовая активность должна быть неким невычислимым образом связана с поддающимся вычислению процессом, который, как утверждают Хамерофф и его коллеги, имеет место внутри этих самых микроканальцев.

Представляемые мною доказательства указывают на то, что распространенные сегодня в некоторых областях науки взгляды ни в коей мере не способствуют хоть сколько-нибудь научному пониманию человеческого разума. И все же это не означает, что феномен сознания так никогда и не найдет своего научного объяснения. Я глубоко убежден — и в этом отношении мои взгляды со времен НРК ничуть не изменились — в том, что научный путь к пониманию феномена разума несомненно существует, и начинаться этот путь должен с более глубокого познания природы собственно физической реальности. Я полагаю чрезвычайно важным, чтобы любой серьезный читатель, намеренный разобраться в том, каким образом столь выдающийся феномен, как разум, может быть объяснен в понятиях материального физического мира, составил бы себе прежде достаточно четкое представление о том, какими странными могут оказаться законы, *в действительности* управляющие этим самым «материалом», из которого состоит наш физический мир.

В конечном счете, именно ради понимания мы и затеяли всю науку, а наука — это все же нечто большее, нежели просто бездумное вычисление.

*Оксфорд,*
*апрель 1994*

*Р. П.*

# БЛАГОДАРНОСТИ

За помощь, оказанную мне в написании этой книги, я весьма обязан многим людям — слишком многим, чтобы поблагодарить каждого из них в отдельности, даже если бы я смог вспомнить все имена. Тем не менее, особую благодарность я хотел бы выразить Гвидо Баччагалуппи и Джереми Баттерфилду за критические замечания, которые они сделали в отношении некоторых частей чернового варианта книги, обнаружив, в частности, серьезную ошибку в моем тогдашнем рассуждении (исправленный текст вошел в третью главу окончательного варианта книги). Кроме того, я благодарен Дэну Айзексону, Абхею Аштекару, Мэри Белл, Брайану Берчу, Джеффу Брукеру, Сьюзан Гринфилд, Робину Гэнди, Роджеру Джеймсу, Дэвиду Дойчу, Эцио Инсинне, Рихарду Йоже, Фрэнсису Крику, Джону Лукасу, Биллу Макколлу, Грэму Мичисону, Клаусу Мозеру, Теду Ньюмену, Джонатану Пенроузу, Оливеру Пенроузу, Стэнли Розену, Рэю Саксу, Грэму Сигалу, Аарону Сломену, Ли Смолину, Рэю Стритеру, Валери Уиллоуби, Соломону Феферману, Эндрю Ходжесу, Дипанкару Хоуму, Дэвиду Чалмерсу, Антону Цайлингеру и в особенности Артуру Экерту за всевозможную информацию и помощь. После выхода в свет моей предыдущей книги (*«Новый разум короля»*) я получил множество устных и письменных отзывов о ней. Пользуясь случаем, хочу поблагодарить всех, кто выразил свое мнение, — оно не пропало даром, хотя на большую часть писем я так и не собрался ответить. Если бы я не извлек пользы из всех этих очень разных комментариев по поводу моей предыдущей книги, вряд ли я ввязался бы в столь устрашающее предприятие, как написание следующей.

Я благодарен организаторам Мессенджеровских лекций в Корнеллском университете (название этого курса лекций совпадает с названием последней главы настоящей книги), Гиффордовских лекций в университете Св. Андрея, Фордеровских лекций в



Новой Зеландии, Грегиногговских лекций в университете Аберистуита и знаменитой серии лекций в Пяти Колледжах (Амхерст, штат Массачусетс), а также многочисленных «разовых» лекций, которые я читал в разных странах. Благодаря этому я получил возможность изложить свои взгляды перед широкой аудиторией и получить ценный отклик. Я благодарен Институту Исаака Ньютона в Кембридже, Сиракузскому университету и университету штата Пенсильвания за их радушие и за присуждение мне званий, соответственно, Почетного внештатного профессора математики и физики, а также Почетного профессора математики и физики Фонда Фрэнсиса и Хелен Пентц. Я также благодарен Национальному научному фонду за поддержку в виде грантов PHY 86-12424 и PHY 43-96246.

Есть, наконец, еще три человека, которые заслуживают особого упоминания. Невозможно переоценить бескорыстную помощь и поддержку, которую оказал мне Энгус Макинтайр, проверив мои рассуждения относительно математической логики в главах 2 и 3 и предоставив мне множество полезной литературы. Выражаю ему свою глубочайшую благодарность. Стюарт Хамерофф рассказал мне о цитоскелете и его микроканальцах; два года назад я и не подозревал о существовании подобных структур! Я очень ему благодарен за эту бесценную информацию, а также за помощь, которую он оказал мне, проверив бóльшую часть материала главы 7. Я навеки у него в долгу за то, что он открыл моим глазам чудеса нового мира. Он, равно как и все остальные, кого я здесь благодарю, конечно же, ни в коей мере не ответственен за те ошибки, совсем избавиться от которых нам так и не удалось. Особо признателен я своей любимой Ванессе по нескольким причинам: за то, что она объяснила мне, почему отдельные части этой книги нужно переписать; за помощь с литературой, что просто спасло меня, а также за ее любовь, терпение и понимание, особенно если учесть, что я постоянно недооцениваю то количество времени, которое отнимает у меня написание книги! Ах, да, чуть не забыл: еще я благодарен ей за то — она, кстати, об этом ничего не знала, — что она отчасти послужила моделью для *вымышленного* образа Джессики, героини придуманной мною истории. Мне очень жаль, что я совсем не знал Ванессу, когда ей было столько же лет, сколько Джессике!

## Источники иллюстраций

Издатели также выражают благодарность правообладателям за разрешение воспроизвести нижеперечисленные иллюстративные материалы.

**Часть I**

Рис. 1.1 A. Nieman/Science Photo Library.

**Часть II**

Рис. 4.12 J. C. Mather *et al.* (1990), *Astrophys. J.*, **354**, L37.

Рис. 5.7 A. Aspect, P. Grangier (1986), *Quantum concepts in space and time* (ed. R. Penrose, C. J. Isham), pp. 1–27, Oxford University Press.

Рис. 5.8 Ashmolean Museum, Oxford.

Рис. 7.2 R. Wichterman (1986), *The biology of paramecium*, 2nd edn., Plenum Press, New York.

Рис. 7.6 Eric Grave/Science Photo Library.

Рис. 7.7 H. Weyl (1943), *Symmetry*, ©1952 Princeton University Press.

Рис. 7.10 N. Hirokawa (1991), *The neuronal cytoskeleton* (ed. R. D. Burgoyne), pp. 5–74, Wiley-Liss, New York.

# ЧИТАТЕЛЮ

Отдельные части этой книги очень сильно отличаются друг от друга в плане использования специальной терминологии. Наиболее специальными являются Приложения A и C, однако бóльшая часть читателей не много потеряет, даже если просто-напросто пропустит все приложения. То же самое можно сказать и о наиболее специальных параграфах второй и, конечно же, третьей главы. Они предназначены, главным образом, для тех читателей, которых нужно убедить в весомости доводов, приводимых мной против чисто вычислительной модели феномена понимания. С другой стороны, менее упорный (или более торопливый) читатель, возможно, предпочтет относительно безболезненный путь к самой сути моего доказательства. Этот путь сводится к прочтению фантастического диалога в § 3.23, предпочтительно предваренному ознакомлением с главой 1, а также с §§ 2.1–2.5 и § 3.1.

С некоторыми вопросами из области более серьезной математики мы встретимся при обсуждении квантовой механики. Речь идет об описаниях гильбертова пространства в §§ 5.12–5.18 и, в особенности, о рассмотрении матрицы плотности в §§ 6.4–6.6, поскольку они весьма важны для понимания того, почему нам, в конечном счете, необходима *более совершенная* теория квантовой механики. Я бы посоветовал читателям, не имеющим математической подготовки (да и тем, кто ее имеет, если уж на то пошлó), при встрече с математическим выражением особенно обескураживающего вида попросту пропускать его, коль скоро станет ясно, что дальнейшее его изучение не приведет к более глубокому пониманию. Тонкости квантовой механики действительно невозможно полностью оценить без некоторого знакомства с ее изящными, но загадочными математическими основами; и все же читатель, без сомнения, уловит какую-то часть присущего ей букета, даже если полностью проигнорирует весь ее математический аппарат.

Кроме того, я должен принести свои извинения читателю еще по одному вопросу. Я вполне способен понять, что моей собеседнице либо собеседнику может не понравиться, вздумай я обратиться к ней или к нему таким образом, который недвусмысленно давал бы понять, что я склонен составлять для себя какое-то мнение относительно ее или его личности, основываясь исключительно на ее или его половой принадлежности, — я, разумеется, никогда так не поступаю! И все же в рассуждениях того сорта, который чаще других встречается в настоящей книге, мне, возможно, придется ссылаться на некую *абстрактную* личность, например, на «наблюдателя» или на «физика». Ясно, что пол этой личности не имеет к теме разговора абсолютно никакого отношения, но в английском языке, к сожалению, нет нейтрального местоимения третьего лица единственного числа. Постоянное же повторение сочетаний типа «он или она» выглядит, безусловно, нелепо. Более того, современная тенденция употреблять местоимения «они», «им» или «их» в качестве местоимений единственного числа в корне неверна грамматически; равным образом я не могу усмотреть ничего хорошего — ни в грамматическом, ни в стилистическом, ни в общечеловеческом плане — в чередовании местоимений «она» и «он», когда речь идет о безличных или метафорических индивидуумах.

Соответственно, в этой книге я избрал политику повсеместного употребления в отношении той или иной абстрактной личности местоимений «он», «ему» или «его». Из этого *ни в коем случае не следует* делать вывода о половой принадлежности упомянутой личности. Эту личность не нужно считать ни мужчиной, ни женщиной. Как правило, индивидуум, которого я называю «он», обладает сознанием и чувствами, а потому называть его «оно»[2], по-моему, не годится. Я искренне надеюсь, что ни одна из моих читательниц не усмотрит личного оскорбления в том, что, говоря в § 5.3, § 5.18 и § 7.12 о своем трехглазом коллеге с $\alpha$-Центавры (абстрактном, разумеется), я использую местоимение «он» и что это же местоимение я употребляю в отношении совершенно безличных индивидуумов в § 1.15, § 4.4, § 6.5, § 6.6 и § 7.10. Я также надеюсь, что ни один из моих читателей не будет обижен тем, что я использую местоимение «она» в отношении умной паучихи

[2]В оригинале «it» — местоимение третьего лица единственного числа, которым в английском языке называют животных и неодушевленные предметы, независимо от их пола и/или рода. — *Прим. перев.*



из § 7.7 и преданной чуткой слонихи из § 8.6 (хотя бы по той простой причине, что в этом случае из контекста очевидно, что обе они *действительно* относятся к женскому полу), а также в отношении демонстрирующей сложное поведение парамеции из § 7.4 (которую я отношу к «женскому» роду по не совсем удовлетворительной причине ее прямой способности к воспроизведению себе подобных), ну и самой матушки-Природы в § 7.7.

Наконец, следует отметить, что ссылки на страницы «Нового разума короля» (НРК) всегда относятся к оригинальному изданию этой книги в твердой обложке. Нумерация страниц американского издания книги в мягкой обложке (Penguin) практически совпадает с оригинальным, а неамериканского издания в мягкой обложке (Vintage) — нет, поэтому номер страницы в последнем можно приблизительно вычислить с помощью формулы:

$$\frac{22}{17} \times n,$$

где $n$ — номер страницы книги в твердой обложке, приводимый здесь в качестве ссылки.

# ПРОЛОГ

Джессика всегда немного нервничала, входя в эту часть пещеры.

— Пап, а что, если тот огромный валун, зажатый между других камней, упадет? Он ведь может загородить выход, и мы уже никогда-никогда не вернемся домой?!

— Он мог бы загородить выход, но этого не случится, — ответил ее отец рассеянно и немного резко, поскольку его, видимо, гораздо больше волновало, как приспосабливаются к сырости и темноте в этом самом дальнем углу пещеры посаженные им растения.

— Но откуда же ты можешь знать, что этого не случится? — упорствовала Джессика.

— Этот валун, вероятно, находится на своем месте уже много тысяч лет и вряд ли упадет именно тогда, когда здесь находимся мы.

Джессику это нисколько не успокоило.

— Все равно он когда-нибудь упадет. Значит, чем дольше он здесь висит, тем больше вероятность того, что он упадет прямо сейчас.

Отец отвлекся от своих растений и, чуть улыбнувшись, посмотрел на Джессику.

— Вовсе нет, — теперь его улыбка стала более заметной, но на лице появилось задумчивое выражение. — Можно даже сказать, что чем дольше он здесь висит, тем *меньше* вероятность его падения при нас. — Дальнейшего объяснения не последовало: отец снова вернулся к своим растениям.

Джессика ненавидела отца, когда у него бывало такое настроение. Хотя — нет: она всегда любила его, любила больше всего и больше всех, но всегда хотела, чтобы он никогда не становился таким, как сейчас. Она знала, что это настроение каким-то образом связано с тем, что он ученый, но до сих пор не понимала каким именно. Она даже надеялась, что сама когда-нибудь



сможет стать ученым, хотя уж она-то позаботится о том, чтобы не впадать в такое состояние духа.

По крайней мере, она перестала беспокоиться, что валун может упасть и загородить вход в пещеру. Она видела, что отец этого не боится, и его уверенность ее успокоила. Она не поняла папиных объяснений, но знала, что в таких случаях он всегда прав — ну или *почти* всегда. Был как-то случай, когда мама с папой поспорили о времени в Новой Зеландии, и мама сказала одно, а папа — совершенно другое. Через три часа папа спустился из своего кабинета, извинился и сказал, что он ошибался, а мама была права. Вид у него при этом был презабавный! «Держу пари, мама тоже могла бы стать ученым, если бы захотела, — подумала про себя Джессика, — и у нее не было бы таких причуд, как у папы».

Следующий вопрос Джессика задала более осторожно, выбрав для этого подходящий момент: отец уже закончил то, чем был занят все это время, но еще не успел начать то, что собирался сделать дальше:

— Пап, я знаю, что валун не упадет. Но давай представим, что он все-таки упал, и нам придется остаться здесь на всю жизнь. В пещере, наверное, станет очень темно. А дышать мы сможем?

— Ну что за глупости! — ответил отец. Затем он прикинул форму и размер валуна и посмотрел на выход из пещеры. — Хм, да-а... похоже, валун достаточно плотно закрыл бы проход. Но воздух все равно проходил бы через оставшиеся щели, так что мы не задохнулись бы. Что касается света, то, я думаю, наверху осталась бы узкая щель, через которую к нам попадал бы свет. Хотя все равно в пещере стало бы очень темно — гораздо темнее, чем сейчас. Но я уверен, что мы смогли бы хорошо видеть, как только привыкли бы к новому освещению. Боюсь, не слишком приятная перспектива! Однако вот что я тебе скажу: если бы мне пришлось провести здесь остаток жизни, то из всех людей на Земле я предпочел бы оказаться здесь со своей замечательной Джессикой и, конечно же, с ее мамой.

Джессика вдруг вспомнила, почему так сильно любит папу.

— Да, для следующего вопроса мне нужна здесь мама: допустим, что валун упал еще до моего рождения, и я появилась у вас здесь, в пещере. Я бы росла вместе с вами прямо тут... а чтобы не умереть от голода, мы могли бы есть твои странные растения.

Отец немного удивленно посмотрел на нее, но промолчал.

— Тогда я не знала бы ничего, *кроме* пещеры. Откуда я могла бы узнать, на что похож реальный мир снаружи? Разве мне пришло бы в голову, что там есть деревья, птицы, кролики и все такое прочее? Конечно, вы могли бы мне о них *рассказать*, ведь вы-то их видели до того, как оказались в пещере. Но как могла бы узнать об этом *я* — именно узнать по-настоящему, *сама*, а не просто поверить в то, что сказали вы?

Ее отец остановился и на несколько минут погрузился в свои мысли. Затем он сказал:

— Ну, думаю, что как-нибудь в солнечный денек какая-нибудь птица могла бы пролететь мимо нашей щели, тогда мы смогли бы увидеть ее тень на стене пещеры. Конечно, ее форма была бы несколько искажена, потому что стена здесь имеет довольно-таки неровную поверхность, но мы смогли бы определить, какую поправку нужно в этом случае сделать. Если бы щель была достаточно узкой и прямой, то птица отбросила бы четкую тень, а если нет, нам пришлось бы вносить и другие поправки. Если бы мимо много раз пролетала бы одна и та же птица, то по ее тени мы смогли бы получить достаточно ясное представление о том, как она на самом деле выглядит, как летает и т. п. Опять же, когда солнце стояло бы низко, а между ним и нашей щелью оказалось бы какое-нибудь дерево с колышущейся кроной, то по его тени мы смогли бы узнать, как оно выглядит. Или мимо щели пробежал бы кролик, и тогда по его тени мы поняли бы, как он выглядит.

— Интересно, — одобрила Джессика. Помолчав немного, она снова спросила:

— А смогли бы мы, если бы застряли здесь, сделать настоящее научное открытие? Представь, что мы сделали большое открытие и устроили здесь большую конференцию — ну, такую же, как те, на которые ты все время ездишь, — чтобы убедить всех, что мы правы. Конечно, все остальные на этой конференции должны, как и мы, прожить в этой пещере всю жизнь, иначе это будет нечестно. Они ведь тоже могут вырасти тут, потому что у тебя очень много разных растений, на *всех* хватит.

На сей раз отец Джессики заметно нахмурился, но снова промолчал. Несколько минут он пребывал в раздумье, затем произнес:



— Да, думаю, такое возможно. Но, видишь ли, самым сложным в этом случае было бы убедить всех, что мир снаружи вообще существует. Все, что они знали бы, — это тени: как они двигаются и как меняются время от времени. Для них сложные извивающиеся тени и фигурки на стене были бы всем, что существует в мире. Поэтому прежде всего нам пришлось бы убедить людей в *существовании* внешнего мира, который описывает наша теория. Собственно говоря, две эти вещи неразрывно связаны. Наличие хорошей теории внешнего мира может стать важным шагом на пути осознания людьми его реального существования.

— Отлично, папа, и какая у нас теория?

— Не так быстро... минуточку... вот: Земля вертится вокруг Солнца!

— Тоже мне *новая* теория!

— Совсем не новая; этой теории, вообще говоря, уже около двадцати трех веков отроду — примерно столько же времени и наш валун висит над входом в пещеру. Но мы же с тобой вообразили, что мы всю жизнь живем в пещере и никто об этом раньше ничего не слыхал. Поэтому нам пришлось бы сначала убедить всех в том, что существуют такие *вещи*, как Солнце, да и сама Земля. Идея же заключается в том, что одна только изящность нашей теории, объясняющей мельчайшие нюансы движения света и тени, в конечном счете убедила бы большинство присутствующих на конференции в том, что эта яркая штука снаружи, которую мы зовем «Солнце», не просто существует, но и что Земля непрерывно движется вокруг нее и при этом еще и вращается вокруг собственной оси.

— А сложно было бы их убедить?

— Очень! Собственно, нам пришлось бы делать два разных дела. Во-первых, нужно было бы показать, каким образом наша простая теория очень точно объясняет огромное количество наиподробнейших данных о том, как движутся по стене яркое пятно и тени, отбрасываемые освещенными им предметами. Это убедило бы некоторых, но нашлись бы и такие, кто указал бы на то, что существует гораздо более «здравая» теория, согласно которой Солнце движется вокруг Земли. При ближайшем рассмотрении эта теория оказалась бы намного сложнее нашей. Но эти люди придерживались бы своей сложной теории — что, вообще говоря, достаточно разумно с их стороны, — поскольку они попросту не смогли бы принять возможности движения их пещеры

со скоростью сто тысяч километров в час, как того требует наша теория.

— Ух ты, а это *на самом деле* правда?

— В некотором роде. Однако во второй части доказательства нам пришлось бы полностью сменить курс и заняться вещами, которые большинство присутствующих на конференции сочли бы совершенно к делу не относящимися. Мы катали бы мячи, раскачивали бы маятники и так далее в том же духе — и все только для того, чтобы показать, что законы физики, управляющие поведением объектов в пещере, ничуть не изменились бы, если бы все содержимое пещеры двигалось в любом направлении с любой скоростью. Этим мы доказали бы, что при движении пещеры с огромной скоростью люди внутри нее и в самом деле никак этого движения не ощутят. Эту очень важную истину пытался доказать еще Галилей. Помнишь, я давал тебе книгу про него?

— Конечно, помню! Боже мой, как все это сложно звучит! Держу пари, что большинство людей на нашей конференции просто уснут — я видела, как они спят на настоящих конференциях, когда ты делаешь доклад.

Отец Джессики едва заметно покраснел:

— Пожалуй, ты права! Но, боюсь, такова наука: куча деталей, многие из которых кажутся скучными и порой совсем не относящимися к делу, даже если заключительная картина оказывается поразительно простой, как и в нашем случае с вращением Земли вокруг своей оси одновременно с ее движением вокруг шарика, называемого Солнцем. Некоторые люди просто не желают вдаваться в подробности, так как находят эту идею достаточно правдоподобной. Но настоящие скептики желают проверить все, выискивая всевозможные слабинки.

— Спасибо, папочка! Так здорово, когда ты рассказываешь мне все это и иногда краснеешь и волнуешься, но, может, мы уже пойдем домой? Темнеет, а я устала и хочу есть. К тому же становится прохладно.

— Ну, пойдем, — отец Джессики накинул ей на плечи свою куртку, собрал вещи и обнял ее, чтобы вывести через уже темнеющий вход. Когда они выходили из пещеры, Джессика еще раз взглянула на валун.

— Знаешь что? Я согласна с тобой, папа. Этот валун запросто провисит здесь еще двадцать три века и даже *дольше*!

# Часть I

# ПОЧЕМУ ДЛЯ ПОНИМАНИЯ РАЗУМА НЕОБХОДИМА НОВАЯ ФИЗИКА?

## Невычислимость сознательного мышления

# 1

# СОЗНАНИЕ И ВЫЧИСЛЕНИЕ

## 1.1. Разум и наука

Насколько широки доступные науке пределы? Подвластны ли ее методам лишь *материальные* свойства нашей Вселенной, тогда как познанию нашей *духовной* сущности суждено навеки остаться за рамками ее возможностей? Или, быть может, однажды мы обретем надлежащее научное понимание тайны разума? Лежит ли феномен сознания человека за пределами досягаемости научного поиска, или все же настанет тот день, когда силой научного метода будет разрешена проблема самого существования наших сознательных «я»?

Кое-кто склонен верить, что мы действительно способны приблизиться к научному пониманию сознания, что в этом феномене вообще нет *ничего* загадочного, а всеми существенными его ингредиентами мы уже располагаем. Они утверждают, что в настоящий момент наше понимание мыслительных процессов человека ограничено лишь крайней сложностью и изощренной организацией человеческого мозга; разумеется, эту сложность и изощренность недооценивать ни в коем случае не следует, однако принципиальных препятствий для выхода за рамки современной научной картины нет. На противоположном конце шкалы расположились те, кто считает, что мы не можем даже надеяться на адекватное применение холодных вычислительных методов бесчувственной науки к тому, что связано с разумом, духом да и самой тайной сознания человека.

В этой книге я попытаюсь обратиться к вопросу сознания с научных позиций. При этом, однако, я твердо убежден (и основано это убеждение на *строго* научной аргументации) в том,

что в современной научной картине мира отсутствует один очень важный ингредиент. Этот недостающий ингредиент совершенно необходим, если мы намерены хоть сколько-нибудь успешно уместить центральные проблемы мыслительных процессов человека в рамки логически последовательного научного мировоззрения. Я утверждаю, что сам по себе этот ингредиент не находится *за пределами*, доступными науке, хотя в данном случае нам, несомненно, придется в некоторой степени расширить наш научный кругозор. Во второй части книги я попытаюсь указать читателю конкретное направление, следуя которому, он непременно придет как раз к такому расширению современной картины физической вселенной. Это направление связано с серьезным изменением самых основных из наших физических законов, причем я весьма детально опишу необходимую природу этого изменения и возможности его применения к биологии нашего мозга. Даже обладая нынешним ограниченным пониманием природы этого недостающего ингредиента, мы вполне способны указать области, отмеченные его несомненным влиянием, и определить, каким именно образом он вносит чрезвычайно существенный вклад в то, что лежит в основе осознаваемых нами ощущений и действий.

Разумеется, некоторые из приводимых мной аргументов окажутся не совсем просты, однако я постарался сделать свое изложение максимально ясным и везде, где только возможно, использовал лишь элементарные понятия. Кое-где в книге все же встречаются некоторые сугубо математические тонкости, но только тогда, когда они действительно необходимы или каким-то образом способствуют достижению более высокой степени ясности рассуждения. С некоторых пор я уже не жду, что смогу с помощью аргументов, подобных приводимым ниже, убедить в своей правоте всех и каждого, однако хотелось бы отметить, что эти аргументы все же заслуживают внимательного и беспристрастного рассмотрения — хотя бы потому, что они создают прецедент, пренебрегать которым нельзя.

Научное мировоззрение, которое на глубинном уровне не желает иметь ничего общего с проблемой сознательного мышления, не может всерьез претендовать на абсолютную завершенность. Сознание является частью нашей Вселенной, а потому любая физическая теория, которая не отводит ему должного места, заведомо неспособна дать истинное описание мира. Я склонен думать, что пока ни одна физическая, биологическая либо математическая теория не приблизилась к объяснению нашего сознания и его логического следствия — интеллекта, однако этот факт ни в коей мере не должен отпугнуть нас от поисков такой теории. Именно эти соображения легли в основу представленных в книге рассуждений. Возможно, продолжая поиски, мы когда-нибудь получим в полной мере приемлемую совокупность идей. Если это произойдет, то наше философское восприятие мира претерпит, по всей вероятности, глубочайшую перемену. И все же научное знание — это палка о двух концах. Важно еще, что мы намерены *делать* со своим научным знанием. Попробуем разобраться, куда могут привести нас наши взгляды на науку и разум.



## 1.2. Спасут ли роботы этот безумный мир?

Открывая газету или включая телевизор, мы всякий раз рискуем столкнуться с очередным проявлением человеческой глупости. Целые страны или отдельные их области пребывают в вечной конфронтации, которая время от времени перерастает в отвратительнейшие войны. Чрезмерный религиозный пыл, национализм, интересы различных этнических групп, просто языковые или культурные различия, а то и корыстные интересы отдельных демагогов могут привести к непрекращающимся беспорядкам и вспышкам насилия, порой беспрецедентным по своей жестокости. В некоторых странах власть до сих пор принадлежит деспотическим авторитарным режимам, которые угнетают народ, держа его под контролем с помощью пыток и бригад смерти. При этом порабощенные — то есть те, кто, на первый взгляд, должны быть объединены общей целью, — зачастую сами конфликтуют друг с другом; создается впечатление, что, получи они свободу, в которой им так долго отказывали, дело может дойти до самого настоящего взаимоистребления. Даже в сравнительно благополучных странах, наслаждающихся преуспеянием, миром и демократическими свободами, природные богатства и людские ресурсы проматываются очевидно бессмысленным образом. Не явный ли это признак общей глупости Человека? Мы уверены, что являем собой апофеоз интеллекта в царстве животных, однако этот интеллект, по всей видимости, оказывается самым жалким образом не способен справиться с множеством проблем, которые продолжает ставить перед нами наше собственное общество.

Впрочем, нельзя забывать и о положительных достижениях нашего интеллекта. Среди них — весьма впечатляющие наука и технология. В самом деле, признавая, что некоторые плоды этой технологии имеют явно спорную долговременную (или сиюминутную) ценность, о чем свидетельствуют многочисленные проблемы, связанные с окружающей средой, и неподдельный ужас перед техногенной глобальной катастрофой, нельзя забывать и о том, что эта же технология является фундаментом нашего современного общества со всеми его удобствами, свободой от страха, болезней и нищеты, с обширными возможностями для интеллектуального и эстетического развития, включая весьма способствующие этому развитию средства глобальной коммуникации. Если технология сумела раскрыть столь огромный потенциал и, в некотором смысле, расширила границы и увеличила возможности наших индивидуальных физических «я», то не следует ли ожидать от нее еще большего в будущем?

Благодаря технологиям — как древним, так и современным — существенно расширились возможности наших органов чувств. Зрение получило поддержку и дополнительную функциональность за счет очков, зеркал, телескопов, всевозможных микроскопов, а также видеокамер, телевизоров и т. п. Не остались в стороне и наши уши: когда-то им помогали слуховые трубки, теперь — крохотные электронные слуховые аппараты; что касается функциональных возможностей нашего слуха, то их расширение связано с появлением телефонов, радиосвязи и спутников. На подмогу естественным средствам передвижения приходят велосипеды, поезда, автомобили, корабли и самолеты. Помощниками нашей памяти выступают печатные книги и фильмы, а также огромные емкости запоминающих устройств *электронных компьютеров*. Наши способности к решению вычислительных задач — простых и рутинных или же громоздких и изощренных — также весьма увеличиваются благодаря возможностям современных компьютеров. Таким образом, технология не только обеспечивает громадное расширение сферы деятельности наших *физических* «я», но и усиливает наши *умственные* возможности, совершенствуя наши способности к выполнению многих повседневных задач. А как насчет тех умственных задач, которые далеки от обыденности и рутины, — задач, требующих участия подлинного *интеллекта*? Совершенно естественно спросить: поможет ли



нам и в их решении технология, основанная на повсеместной компьютеризации?

Я практически не сомневаюсь, что в нашем технологическом (часто сплошь компьютеризованном) обществе в неявном виде присутствует, как минимум, одно направление, содержащее громадный потенциал для совершенствования интеллекта. Я имею в виду образовательные возможности нашего общества, которые могли бы весьма значительно выиграть от применения различных аспектов технологии, — для этого требуются лишь должные чуткость и понимание. Технология обеспечивает необходимый потенциал, т. е. хорошие книги, фильмы, телевизионные программы и всевозможные интерактивные системы, управляемые компьютерами. Эти и прочие разработки предоставляют массу возможностей для расширения нашего кругозора; они же, впрочем, могут и задушить его. Человеческий разум способен на гораздо большее, чем ему обычно дают шанс достичь. К сожалению, эти возможности зачастую попросту разбазариваются, и умы как старых, так и малых не получают тех благоприятных возможностей, которых они несомненно заслуживают.

Многие читатели спросят: а нет ли какой-то иной возможности существенного расширения умственных способностей человека — например, с помощью этакого нечеловеческого электронного «интеллекта», к появлению которого нас как раз вплотную подводят выдающиеся достижения компьютерных технологий? Действительно, уже сейчас мы часто обращаемся за интеллектуальной поддержкой к компьютерам. В очень многих ситуациях человек, используя лишь свой невооруженный разум, оказывается не в состоянии оценить возможные последствия того или иного своего действия, так как они могут находиться далеко за пределами его ограниченных вычислительных способностей. Таким образом, можно ожидать, что в будущем произойдет значительное расширение роли компьютеров именно в этом направлении, т. е. там, где для принятия решения человеческому интеллекту требуются именно однозначные и вычислимые факты.

И все же не могут ли компьютеры достичь в конечном итоге чего-то большего? Многие специалисты заявляют, что компьютеры обладают потенциалом, достаточным — по крайней мере, принципиально — для формирования *искусственного* интеллекта, который со временем превзойдет наш собственный[1]. По утверждению этих специалистов, как только управляемые по-

средством вычислительных схем роботы достигнут уровня «эквивалентности человеку», понадобится совсем немного времени, чтобы они значительно поднялись над нашим ничтожным уровнем. Только *тогда*, не унимаются специалисты, появятся у нас власти, обладающие интеллектом, мудростью и пониманием, достаточными для того, чтобы суметь разрешить глобальные проблемы этого мира, человечеством же и созданные.

Когда же нам следует ожидать наступления сего счастливого момента? По данному вопросу у упомянутых специалистов нет единого мнения. Одни говорят о многих столетиях, другие заявляют, будто эквивалентность компьютера человеку будет достигнута всего через несколько десятилетий(2). Последние обычно указывают на очень быстрый «экспоненциальный» рост мощности компьютеров и основывают свои оценки на сравнении скорости и точности транзисторов с относительной медлительностью и «небрежностью» нейронов. И правда, скорость работы электронных схем уже более чем в миллион раз превышает скорость возбуждения нейронов в мозге (порядка $10^9$ операций в секунду для транзисторов и лишь $10^3$ для нейронов[1]), при этом электронные схемы демонстрируют высокую точность синхронизации и обработки инструкций, что ни в коей мере не свойственно нейронам. Более того, конструкции «принципиальных схем» мозга присуща высокая степень случайности, что, на первый взгляд, представляется весьма серьезным недостатком по сравнению с продуманной и точной организацией электронных печатных плат.

Кое в чем, однако, нейронная структура мозга все же вполне измеримо превосходит современные компьютеры, хотя это превосходство может оказаться относительно недолговечным. Ученые утверждают, что по общему количеству нейронов (несколько сотен тысяч миллионов) человеческий мозг опережает — в пересчете на транзисторы — современные компьютеры. Более того, в среднем, нейроны мозга соединены гораздо бóльшим количеством *связей*, нежели транзисторы в компьютере. В частности, клетки Пуркинье в мозжечке могут иметь до 80 000 синаптических окончаний (зон контакта между нейронами), тогда как для компьютера соответствующее значение равно максимум трем или четырем. (В дальнейшем я приведу еще несколько комментариев относительно мозжечка; см. § 1.14, § 8.6.) Кроме того, бóльшая часть транзисторов в современных компьютерах занимается лишь хранением данных и не имеет отношения непосредственно к вычислениям, тогда как в мозге, по всей видимости, в вычислениях может принимать участие гораздо более значительный процент клеток.

[1] Микросхема Intel Pentium содержит более трех миллионов транзисторов на «кремниевой пластине» размером с ноготь большого пальца, причем каждый из этих транзисторов способен на 113 миллионов полных циклов в секунду.



Это временное превосходство мозга может быть без труда преодолено в будущем, особенно когда должное развитие получат вычислительные системы с массивным «параллелизмом». Преимущество компьютеров в том, что отдельные их узлы можно объединять друг с другом, создавая все более крупные блоки, так что общее количество транзисторов, в принципе, можно увеличивать почти бесконечно. Кроме того, ждут своего выхода на сцену и технологические инновации — такие, как замена кабелей и транзисторов современных компьютеров соответствующими оптическими (лазерными) устройствами, благодаря чему, вероятно, будет достигнуто огромное увеличение скорости и мощности с одновременным уменьшением размеров компьютеров. На более фундаментальном уровне можно отметить, что наш мозг, судя по всему, *застрял* на своем теперешнем уровне, и его количественные характеристики вряд ли в обозримом будущем изменятся; кроме того, имеется и много других ограничений — например, мозг вырастает из одной-единственной клетки, и ничего с этим не поделаешь. Компьютеры же можно конструировать, учитывая заранее возможность их расширения по мере необходимости. Хотя несколько позже я укажу на некоторые важные факторы, которые в данном рассуждении пока не фигурируют (в частности, речь пойдет о весьма бурной деятельности, лежащей в основе функционирования нейронов), одна лишь вычислительная мощь компьютеров вполне способна составить очень и очень внушительный довод в пользу следующего неутешительного предположения: если машина на данный момент и не превосходит человеческий мозг, то она *непременно* превзойдет его в самом ближайшем будущем.

Таким образом, если поверить самым смелым заявлениям наиболее отъявленных провозвестников искусственного интеллекта и допустить, что компьютеры и управляемые ими роботы в конечном счете — и даже, вероятно, довольно скоро — во всем превзойдут человека, то получается, что компьютеры способны стать чем-то неизмеримо большим, чем просто помощниками *на-*

*шего* интеллекта. Они, в сущности, разовьют свой собственный колоссальный интеллект. А мы сможем обращаться к этому высшему интеллекту за советом и поддержкой во всех своих заботах — и наконец-то появится возможность исправить все то зло, что мы принесли в этот мир!

Однако из этих потенциальных соображений возможно, по-видимому, и другое логическое следствие, причем весьма и весьма тревожное. Не сделают ли такие компьютеры в итоге ненужными самих людей? Если управляемые компьютерами роботы превзойдут нас во всех отношениях, то не обнаружат ли они, что машины в состоянии править миром неизмеримо лучше людей, и не сочтут ли они нас в таком случае вообще ни на что не пригодными? Все человечество окажется в таком случае не более чем пережитком прошлого. Быть может, если повезет, они оставят нас при себе в качестве домашних животных, как однажды предположил Эдвард Фредкин. Возможно также, что у нас достанет сообразительности, и мы сумеем перенести «информационные модели», составляющие нашу «сущность», в машинную форму — о такой возможности писал Ханс Моравек (1988). Опять же, может, и *не* повезет, а сообразительности *не* достанет...

## 1.3. Вычисление и сознательное мышление

В чем же здесь загвоздка? Неужели все дело лишь в вычислительных способностях, в скорости и точности работы, в объеме памяти или, быть может, в конкретном способе «связи» отдельных структурных элементов? С другой стороны, не может ли наш мозг выполнять какие-то действия, которые вообще невозможно описать через вычисление? Каким образом можно поместить в такую вычислительную картину нашу способность к осмысленному осознанию — счастья, боли, любви, какого-либо эстетического переживания, желания, понимания и т. п.? Будут ли компьютеры будущего действительно обладать *разумом*? Влияет ли обладание сознательным разумом на поведение индивида, и если влияет, то как именно? Имеет ли вообще смысл говорить о таких вещах на языке научных терминов; иными словами, обладает ли наука достаточной компетентностью для того, чтобы рассматривать вопросы, относящиеся к сознанию человека?

Мне кажется, что можно говорить, как минимум, о четырех различных точках зрения[3] — или даже крайностях, — которых



разумный индивид может придерживаться в отношении данного вопроса:

$\mathscr{A}$. Всякое мышление есть вычисление; в частности, ощущение осмысленного осознания есть не что иное, как результат выполнения соответствующего вычисления.

$\mathscr{B}$. Осознание представляет собой характерное проявление физической активности мозга; хотя любую физическую активность можно моделировать посредством той или иной совокупности вычислений, численное моделирование как таковое не способно вызвать осознание.

$\mathscr{C}$. Осознание является результатом соответствующей физической активности мозга, однако эту физическую активность невозможно должным образом смоделировать вычислительными средствами.

$\mathscr{D}$. Осознание невозможно объяснить в физических, математических и вообще научных терминах.

Точка зрения $\mathscr{D}$, полностью отрицающая взгляды физикалистов и рассматривающая разум как нечто абсолютно неподвластное языку науки, свойственна мистикам; и, по крайней мере, в какой-то степени, такое мировоззрение, видимо, сродни религиозной доктрине. Лично я считаю, что связанные с разумом вопросы, пусть даже и не объясняемые должным образом в рамках современного научного понимания, не следует рассматривать как нечто, чего науке никогда не постичь. Пусть на данный момент наука и не способна сказать в отношении этих вопросов своего веского слова, со временем ее возможности неминуемо расширятся настолько, что в ней найдется место и для таких вопросов, причем не исключено, что в процессе такого расширения изменятся и сами ее методы. Отбрасывая мистицизм с его отрицанием научных критериев в пользу научного познания, я все же убежден, что и в рамках усовершенствованной науки вообще и математики в частности найдется немало загадок, среди которых не последнее место займет тайна разума. К некоторым из этих идей я еще вернусь в следующих главах книги, сейчас же достаточно будет сказать, что согласиться с точкой зрения $\mathscr{D}$ я никак не могу, поскольку твердо намерен двигаться вперед, следуя пути, проложенному наукой. Если мой читатель питает сильное убеждение,

что истинным является именно пункт $\mathscr{D}$, в той или иной его форме, я попрошу его потерпеть еще немного и посмотреть, сколько нам удастся пройти вместе по дороге науки, — и попытаться при этом понять, куда, по моему убеждению, эта дорога в конечном счете нас приведет.

Теперь обратимся к противоположной крайности: к точке зрения $\mathscr{A}$. Эту точку зрения разделяют сторонники так называемого *сильного*, или *жесткого, искусственного интеллекта* (*ИИ*); иногда для обозначения такой позиции употребляется также термин *функционализм*[4], хотя некоторые распространяют термин «функционализм» еще и на определенные варианты пункта $\mathscr{C}$. Одни считают $\mathscr{A}$ единственно возможной точкой зрения, которую допускает сугубо научное отношение. Другие воспринимают $\mathscr{A}$ как нелепость, которая вряд ли сто́ит сколь-нибудь серьезного внимания. Существует, несомненно, множество различных вариантов позиции $\mathscr{A}$. (Длинный список альтернативных версий вычислительной точки зрения приводится в [344].) Некоторые из них отличаются лишь различным пониманием того, что следует считать «вычислением» или «выполнением вычисления». Есть и такие приверженцы $\mathscr{A}$, которые вообще не считают себя «сторонниками сильного ИИ», поскольку придерживаются принципиально иного взгляда на интерпретацию термина «вычисление», нежели та, что предлагается в традиционном понятии ИИ (см. [112]). Я рассмотрю эти вопросы подробнее в § 1.4. Пока же достаточно будет понимать под «вычислением» такую операцию, какую способны выполнять обычные универсальные компьютеры. Другие сторонники позиции $\mathscr{A}$ могут расходиться в интерпретации значения терминов «осмысление» или «осознание». Некоторые отказываются признавать само *существование* такого феномена, как «осмысленное осознание», тогда как другие собственно феномен признают, однако рассматривают его лишь как своего рода «эмергентное свойство» (см. также § 4.3 и § 4.4), которое проявляется всякий раз, когда выполняемое вычисление имеет достаточную степень сложности (или громоздкости, или самоотносимости, или чего угодно еще). В § 1.12 я приведу свою собственную интерпретацию терминов «осознание» и «осмысление». Пока же любые расхождения в возможной их интерпретации не будут иметь особой важности для наших рассуждений.

Аргументы, приведенные мной в НРК, были направлены, главным образом, против точки зрения $\mathscr{A}$, или позиции сильного ИИ. Один только объем этой книги должен показать, что, хотя лично я не верю в истинность $\mathscr{A}$, я все же рассматриваю эту точку зрения как реальную возможность, на которую сто́ит обратить серьезное внимание. $\mathscr{A}$ есть следствие предельно операционного подхода к науке, предполагающего, что абсолютно все феномены физического мира можно описать одними лишь вычислительными методами. В одной из крайних вариаций такого подхода сама Вселенная рассматривается, по существу, как единый гигантский компьютер[5], причем «осмысленные осознания», формирующие, в сущности, наш с вами сознательный разум, вызываются посредством соответствующих субвычислений, выполняемых этим компьютером.



Я полагаю, что эта точка зрения (согласно которой физические системы следует считать простыми вычислительными объектами) отчасти основывается на значительной и постоянно растущей роли вычислительных моделей в современной науке и отчасти из убеждения в том, что сами физические объекты — это, в некотором смысле, всего лишь «информационные модели», подчиняющиеся математическим, вычислительным законам. Бóльшая часть материи, из которой состоят наше тело и мозг, постоянно обновляется — неизменными остаются лишь их *модели*. Более того, и сама материя, судя по всему, ведет преходящее существование, поскольку ее можно преобразовать из одной формы в другую. Даже *масса* материального тела, которая является точной физической мерой количества материи, содержащегося в теле, может быть при определенных обстоятельствах превращена в чистую энергию (в соответствии со знаменитой формулой Эйнштейна $E = mc^2$). Следовательно, и материальная субстанция, по-видимому, способна превращаться в нечто, обладающее лишь теоретико-математической реальностью. Более того, если верить квантовой теории, материальные частицы — это не что иное, как информационные «волны». (На этих вопросах мы более подробно остановимся во второй части книги.) Таким образом, сама материя есть нечто неопределенное и недолговечное, поэтому вполне разумно предположить, что постоянство человеческого «я», возможно, больше связано с сохранением *моделей*, нежели реальных частиц материи.

Даже если мы не считаем возможным рассматривать Вселенную всего лишь как компьютер, к точке зрения $\mathscr{A}$ нас могут подтолкнуть более практические, операционные соображения.

Предположим, что перед нами управляемый компьютером робот, который отвечает на вопросы так же, как это делал бы человек. Мы спрашиваем его, как он себя чувствует, и обнаруживаем, что его ответы полностью соответствуют нашим представлениям об ответах на подобные вопросы разумного существа, действительно обладающего чувствами. Он говорит нам, что способен к осознанию, что ему весело или грустно, что он воспринимает красный цвет и что его волнуют вопросы «разума» и «собственного я». Он может даже выразить озадаченность: следует ли ему допустить, что и *других* существ (в частности, людей) нужно рассматривать как обладающих сознанием, сходным с тем, на обладание которым претендует он сам. Что помешает нам поверить *его* утверждениям о том, что он ощущает, любопытствует, радуется, испытывает боль, особенно если учесть, что о других людях мы знаем ничуть не больше и все же *считаем* их обладающими сознанием? Мне кажется, что операционный аргумент все же обладает значительной силой, хотя его и нельзя считать решающим. Если все *внешние* проявления сознательного разума, включая ответы на непрекращающиеся вопросы, действительно могут быть полностью воспроизведены системой, управляемой исключительно вычислительными алгоритмами, то мы имеем полное право допустить, что в рамках рассматриваемой ситуации такая модель должна содержать и все *внутренние* проявления разума (включая собственно сознание).

Принимая или отвергая такой вывод из вышеприведенного рассуждения, которое в основе своей составляет суть так называемого *теста Тьюринга*[6], мы тем самым определяем свою принадлежность к тому или иному лагерю — именно здесь проходит граница между позициями $\mathscr{A}$ и $\mathscr{B}$. Согласно $\mathscr{A}$, любого управляемого компьютером робота, который после достаточно большого количества заданных ему вопросов ведет себя так, *словно* он обладает сознанием, следует *фактически* считать обладающим сознанием. Согласно $\mathscr{B}$, робот вполне может вести себя точно так же, как обладающий сознанием человек, при этом реально не имея и малой доли этого внутреннего качества. И $\mathscr{A}$, и $\mathscr{B}$ сходятся в том, что управляемый компьютером робот может *вести себя* так, как ведет себя обладающий сознанием человек. $\mathscr{C}$ же, напротив, не допускает и малейшей возможности того, что когда-либо может быть реализована эффективная модель обладающего сознанием человека в виде управляемого компьютером робота. Таким образом, согласно $\mathscr{C}$, после некоторого достаточно большого количества вопросов реальное отсутствие сознания у робота так или иначе проявится. Вообще говоря, $\mathscr{C}$ является в гораздо большей степени *операционной* точкой зрения, нежели $\mathscr{B}$, и в этом отношении она больше похожа на $\mathscr{A}$, чем на $\mathscr{B}$.



Так что же представляет собой позиция $\mathscr{B}$? Я думаю, что $\mathscr{B}$ — это, вероятно, именно та точка зрения, которую многие полагают «научным здравым смыслом». Описываемый ею искусственный интеллект еще называют *слабым* (или *мягким*) ИИ. Подобно $\mathscr{A}$, она утверждает, что все физические объекты этого мира должны вести себя в соответствии с некоторыми научными положениями, которые, в принципе, допускают создание вычислительной модели этих объектов. С другой стороны, эта точка зрения уверенно отрицает мнение операционистов, согласно которому любой объект, внешне проявляющий себя как сознательное существо, непременно обладает сознанием. Как отмечает философ Джон Серл[7], вычислительную модель физического процесса никоим образом не следует отождествлять с самим процессом, происходящим в действительности. (Компьютерная модель, например, урагана — это совсем не то же самое, что и реальный ураган!) Согласно взгляду $\mathscr{B}$, наличие или отсутствие сознания очень сильно зависит от того, какой именно физический объект «осуществляет мышление» и какие физические действия он при этом совершает. И только потом следует рассмотреть конкретные вычисления, которых требуют эти действия. Таким образом, активность биологического мозга может вызвать осознание, а вот его точная электронная модель вполне может оказаться на это неспособной. Это различие, по $\mathscr{B}$, совсем не обязательно должно оказаться различием между биологией и физикой. Однако крайне важным остается реальное *материальное* строение рассматриваемого объекта (скажем, мозга), а не просто его вычислительная активность.

Позиция $\mathscr{C}$, на мой взгляд, ближе всех к истине. Она подразумевает более операционный подход, нежели $\mathscr{B}$, так как утверждает, что существуют такие внешние проявления обладающих сознанием объектов (скажем, мозга), которые отличаются от внешних проявлений компьютера: внешние проявления сознания невозможно должным образом воспроизвести вычислительными методами. Свои основания для такой убежденности я приведу несколько позже. Поскольку $\mathscr{C}$, как и $\mathscr{B}$, не отвергает позиции

физикалистов, согласно которой разум возникает в результате проявления активности тех или иных физических объектов (например, мозга, хотя это и не обязательно), $\mathscr{C}$ подразумевает, что *не* всякую физическую активность можно должным образом смоделировать вычислительными методами.

Допускает ли современная физика возможность существования процессов, которые принципиально невозможно смоделировать на компьютере? Если мы надеемся получить на этот вопрос математически строгий ответ, то нас ждет разочарование. По крайней мере, лично мне такой ответ неизвестен. Вообще, с математической точностью здесь дело обстоит несколько запутаннее, чем хотелось бы[8]. Однако сам я убежден в том, что подобные невычислимые процессы следует искать *за пределами* тех областей физики, которые описываются известными на настоящий момент физическими законами. Далее в этой книге я вновь перечислю некоторые весьма серьезные — причем именно физические — доводы в пользу того, что мы действительно нуждаемся в новом взгляде на ту область, которая лежит между уровнем микроскопических величин, где господствуют квантовые законы, и уровнем «обычных» размеров, подвластным классической физике. Хотя, надо сказать, далеко не все современные физики единодушно уверены в необходимости подобной новой физической теории.

Таким образом, существуют, как минимум, две различные точки зрения, которые можно отнести к категории $\mathscr{C}$. Одни сторонники $\mathscr{C}$ утверждают, что наше современное физическое понимание абсолютно адекватно, следует лишь обратить в рамках традиционной теории более пристальное внимание на некоторые тонкие типы поведения, которые вполне могут вывести нас за пределы того, что целиком и полностью объяснимо с помощью вычислений (некоторые из таких типов мы рассмотрим ниже — например, хаотическое поведение (§ 1.7), некоторые тонкости непрерывного действия в противоположность дискретному (§ 1.8), квантовая случайность). Другие же, напротив, полагают, что современная физика, в сущности, не располагает должными средствами для реализации невычислимости требуемого типа. Далее я представлю некоторые веские, на мой взгляд, доводы в пользу принятия позиции $\mathscr{C}$ именно в этом, более строгом, ее варианте, который предполагает создание фундаментально новой физики.



Кое-кто попытался было объявить, что эти соображения отправляют меня прямиком в лагерь сторонников точки зрения $\mathscr{D}$, поскольку я утверждаю, что для отыскания хоть какого-то объяснения феномену сознания нам придется выйти за пределы известной науки. Однако между упомянутым строгим вариантом $\mathscr{C}$ и точкой зрения $\mathscr{D}$ есть существенная разница, в частности, на уровне *методологии*. В соответствии с $\mathscr{C}$, проблема осмысленного осознания носит, в сущности, научный характер, даже если подходящей наукой мы пока что не располагаем. Я всецело поддерживаю эту точку зрения; я полагаю, что ответы на интересующие нас вопросы нам следует искать именно с помощью научных методов — разумеется, должным образом усовершенствованных, пусть даже о конкретной природе необходимых изменений мы, возможно, имеем на данный момент лишь самое смутное представление. В этом и состоит ключевая разница между $\mathscr{C}$ и $\mathscr{D}$, насколько бы похожими ни казались нам соответствующие мнения относительно того, на что способна *современная* наука.

Определенные выше точки зрения $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$, $\mathscr{D}$ представляют собою крайности, или полярные точки возможных позиций, которых может придерживаться тот или иной индивидуум. Я вполне допускаю, что кому-то может показаться, что их собственные взгляды не подходят ни под одну из перечисленных категорий, а лежат где-то между ними либо противоречат некоторым из них. Безусловно, между такими, например, крайними точками зрения, как $\mathscr{A}$ и $\mathscr{B}$, можно разместить множество различных промежуточных точек зрения (см. [344]). Существует даже мнение (весьма, кстати, широко распространенное), которое лучше всего определяется как комбинация $\mathscr{A}$ и $\mathscr{D}$ (или, быть может, $\mathscr{B}$ и $\mathscr{D}$), — предусматриваемая им возможность еще сыграет немаловажную роль в наших дальнейших размышлениях. Согласно этому мнению, мозг действительно работает как компьютер, однако компьютер настолько невообразимой сложности, что его имитация не под силу человеческому и научному разумению, ибо он, несомненно, является божественным творением Господа — «лучшего в мире системотехника», не иначе![9]

## 1.4. Физикализм и менталим

Я должен сделать здесь краткое отступление касательно использования терминов «физикалист» и «менталист» (обычно противопоставляемых один другому), в нашей конкретной

ситуации, т. е. в отношении крайних точек зрения, обозначенных нами через $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$ и $\mathscr{D}$. Поскольку $\mathscr{D}$ являет собой полное отрицание физикализма, сторонников $\mathscr{D}$ безусловно следует считать менталистами. Однако мне не совсем ясно, где провести границу между физикализмом и ментализмом в случае с тремя другими позициями $\mathscr{A}$, $\mathscr{B}$ и $\mathscr{C}$. Я полагаю, что приверженцев $\mathscr{A}$ следует обыкновенно считать физикалистами, и я уверен, что подавляющее *их* большинство согласилось бы со мной. Однако здесь скрывается некий парадокс. В соответствии с $\mathscr{A}$, *материальное* строение мыслящего устройства считается несущественным. Все его мыслительные атрибуты определяются лишь вычислениями, которые это устройство выполняет. Сами по себе вычисления суть феномены абстрактной математики, не связанные с конкретными материальными телами. Таким образом, согласно $\mathscr{A}$, сами мыслительные атрибуты не имеют жесткой связи с физическими объектами, а потому термин «физикалист» может показаться несколько неуместным. Точки зрения $\mathscr{B}$ и $\mathscr{C}$, напротив, требуют, чтобы при определении наличия в том или ином объекте подлинного разума решающую роль играло реальное физическое строение рассматриваемого объекта. Соответственно, вполне можно было бы утверждать, что именно эти точки зрения, а никак не $\mathscr{A}$, представляют возможные позиции физикалистов. Однако такая терминология, по-видимому, вошла бы в некоторое противоречие с общепринятым употреблением, где более уместным считается называть «менталистами» сторонников $\mathscr{B}$ и $\mathscr{C}$, поскольку в этих случаях свойства мышления рассматриваются как нечто «реальное», а не просто как «эпифеномены»[2], которые случайным образом возникают при выполнении определенных типов вычислений. Ввиду такой путаницы, я буду избегать использования терминов «физикалист» и «менталист» в последующих рассуждениях, ссылаясь вместо этого на конкретные точки зрения $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$ и $\mathscr{D}$, определенные выше.

## 1.5. Вычисление: нисходящие и восходящие процедуры

До сих пор было не совсем ясно, что именно я понимаю под термином «вычисление» в определениях позиций $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$ и $\mathscr{D}$,

[2]Эпифеномен — побочное явление, сопутствующее другим явлениям (феноменам), но не оказывающее на них никакого влияния. — *Прим. перев.*



приведенных в § 1.3. Что же такое вычисление? В двух словах: это все, что делает самый обычный универсальный компьютер. Если же мы хотим быть более точными, то следует воспринимать этот термин в соответственно идеализированном смысле: *вычисление* — это действие *машины Тьюринга*.

А что такое машина Тьюринга? По сути, это и есть математически идеализированный компьютер (теоретический предшественник современного универсального компьютера); идеализирован же он в том смысле, что никогда не ошибается, может работать сколько угодно долго и обладает неограниченным объемом памяти. Немного более подробно о точных спецификациях машин Тьюринга я расскажу в § 2.1 и в Приложении А (с. 193). (Интересующийся более полным введением в этот вопрос читатель может обратиться к описанию, приведенному в НРК, глава 2, а также к работам Клина [223] или Дэвиса [72].)

Для описания деятельности машины Тьюринга нередко используют термин «алгоритм». В данном контексте я считаю термин «алгоритм» полностью синонимичным термину «вычисление». Здесь необходимо небольшое разъяснение, так как в отношении термина «алгоритм» некоторые придерживаются более узкой точки зрения, нежели предлагаемая мною здесь, подразумевая под алгоритмом то, что я в дальнейшем буду более конкретно называть «нисходящим алгоритмом». Попытаемся разобраться, что же следует понимать в контексте вычисления под термином «нисходящий» и противоположным ему термином «восходящий».

Мы говорим, что вычислительная процедура имеет *нисходящую* организацию, если она построена в соответствии с некоторой прозрачной и хорошо структурированной фиксированной вычислительной процедурой (которая может содержать некий заданный заранее объем данных) и предоставляет, в частности, четкое решение для той или иной рассматриваемой проблемы. (Описанный в НРК на с. 31[3] евклидов алгоритм нахождения наибольшего общего делителя двух натуральных чисел представляет собой простой пример нисходящего алгоритма.) В противоположность такой организации существует организация *восходящая*, где упомянутые четкие правила выполнения действий и объем данных заранее не определены, однако вместо этого имеет-

[3]Напомним, что здесь и далее приводятся страницы оригинального английского издания. — *Прим. перев.*

ся некоторая процедура, определяющая, каким образом система должна «обучаться» и повышать свою эффективность в соответствии с накопленным «опытом». Иными словами, в случае восходящей системы правила выполнения действий подвержены постоянному изменению. Очевидно, что такая система должна пройти множество циклов, выполняя требуемые действия над непрерывно поступающими данными. Во время каждого прогона производится оценка эффективности (возможно, самой системой), после чего, в соответствии с этой оценкой, система так или иначе модифицирует свои действия, стремясь улучшить качество вывода данных. Например, на вход системы подаются несколько оцифрованных с некоторым качеством фотопортретов, и ставится задача — определить, на каких портретах изображен один человек, а на каких — другой. После каждого прогона результат выполнения задачи сравнивается с правильным, после чего правила выполнения действий модифицируются так, чтобы с некоторой вероятностью добиться улучшения функционирования системы при следующем прогоне.

Конкретные способы такого улучшения в какой-либо конкретной восходящей системе нас в данный момент не интересуют. Достаточно сказать, что количество всевозможных готовых схем весьма велико. Среди наиболее известных систем восходящего типа можно упомянуть так называемые *искусственные нейронные сети* (иногда их называют просто «нейронными сетями», что может ввести в некоторое заблуждение), которые представляют собой компьютерные самообучающиеся программы — или же особым образом сконструированные электронные устройства, — основанные на определенных представлениях о реальной организации системы связей между нейронами в мозге и о том, каким образом эта система улучшается по мере приобретения мозгом опыта. (Вопрос о том, как *в действительности* модифицирует самоё себя система взаимосвязей между нейронами мозга, приобретет для нас особую значимость несколько позднее; см. §7.4 и §7.7.) Очевидно также, что возможны системы, сочетающие в себе элементы как восходящей, так и нисходящей организации.

Для наших целей важно понимать, что и нисходящие, и восходящие вычислительные процедуры с легкостью выполняются на универсальном компьютере, а потому их можно отнести к категории процессов, названных мною *вычислительными* и *алгоритмическими*. Таким образом, в случае восходящих (или комбинированных) систем сам *способ* модификации системой своих процедур задается какими-то целиком и полностью вычислительными инструкциями, причем задается заблаговременно. Этим и объясняется возможность реализации всей системы на обычном компьютере. Существенная *разница* между восходящей (или комбинированной) системой и системой нисходящей состоит в том, что в первом случае вычислительная процедура должна подразумевать возможность сохранения «памяти» о предыдущем выполнении задачи (т. е. обладать способностью накапливать «опыт») с тем, чтобы эту память затем можно было использовать в последующих вычислительных действиях. Конкретные подробности сейчас не имеют особого значения, однако к обсуждению этого вопроса мы еще вернемся в §3.11.



Задавшись целью создать *искусственный интеллект* (сокращенно «ИИ»), человек пока лишь пытается сымитировать разумное поведение на каком угодно уровне посредством каких-то вычислительных средств. При этом часто используется как нисходящая, так и восходящая организация. Первоначально наиболее перспективными представлялись нисходящие системы[10], однако сейчас все большую популярность приобретают восходящие системы типа искусственной нейронной сети. По всей видимости, получения наиболее успешных систем ИИ можно ожидать лишь при том или ином *сочетании* нисходящих и восходящих организаций. У каждой из них есть свои преимущества. Нисходящая организация наиболее успешна в тех областях, где данные и правила выполнения действий четко определены и имеют хорошо выраженный вычислительный характер, — при решении некоторых конкретных математических задач, создании вычислительных систем для игры в шахматы или, скажем, в медицинской диагностике, где определение того или иного заболевания происходит с помощью заданных наборов правил, основанных на общепринятых медицинских процедурах. Восходящая же организация оказывается полезной, когда критерии для принятия решений не слишком точны или не совсем ясны, — как, например, при распознавании лиц или звуков или, возможно, при поиске месторождений минералов, где основным поведенческим критерием становится повышение эффективности на основе накопленного опыта. Во многих подобных системах действительно присутству-

ют элементы *и* нисходящей, *и* восходящей организаций (например, шахматный компьютер, обучающийся на основе опыта, или созданное на базе какой-либо четкой геологической теории вычислительное устройство, помогающее в поисках месторождений минералов).

Я думаю, справедливым будет сказать, что лишь в некоторых примерах нисходящей (или по большей части нисходящей) организации компьютеры демонстрируют значительное превосходство над человеком. Самым очевидным примером может служить прямой численный расчет, где в наше время компьютеры побеждают человека без каких-либо усилий. То же самое относится и к «вычислительным» играм, типа шахмат и шашек, в которые у лучших компьютеров способны выиграть, возможно, лишь несколько человек (более подробно об этом в § 1.15 и § 8.2). В случае же восходящей организации (искусственной нейронной сети) компьютерам лишь в немногих специфических примерах удается достичь приблизительно уровня обычных хорошо обученных людей.

Еще одно отличие между видами компьютерных систем связано с различием между *последовательной* и *параллельной* архитектурами. Компьютер последовательного действия — это машина, выполняющая вычисления друг за другом, поэтапно, тогда как параллельный компьютер выполняет множество независимых вычислений одновременно, результаты же этих вычислений сводятся вместе лишь по завершении достаточно большого их количества. Кстати, у истоков разработки некоторых параллельных систем стояли все те же теории, описывающие предполагаемые способы функционирования мозга. Здесь следует отметить, что различие между вычислительными машинами последовательного и параллельного действия ни в коей мере не является *принципиальным*. Параллельное действие всегда можно смоделировать последовательно, хотя, конечно же, существуют некоторые типы задач (весьма немногочисленные), для решения которых эффективнее (в смысле затрат времени на вычисление и т. п.) будет параллельное действие, нежели последовательное. Поскольку в рамках настоящего труда меня занимают, главным образом, принципиальные вопросы, различия между параллельными и последовательными вычислениями не представляются в этом отношении особенно существенными.



## 1.6. Противоречит ли точка зрения $\mathscr{C}$ тезису Черча–Тьюринга?

Вспомним, что точка зрения $\mathscr{C}$ предполагает, что обладающий сознанием мозг функционирует таким образом, что его активность не поддается никакому численному моделированию — ни нисходящего, ни восходящего, ни какого-либо другого типа. Те, кто сомневается в истинности $\mathscr{C}$, могут отчасти оправдать свои сомнения тем, что формулировка $\mathscr{C}$ якобы противоречит так называемому *тезису Черча* (или тезису Черча–Тьюринга) — вернее, тому условию, которое сейчас общепринято обозначать упомянутым термином. В чем же суть тезиса Черча? В первоначальной форме, предложенной американским логиком Алонзо Черчем в 1936 году, этот тезис гласил, что любой процесс, который можно корректно назвать «чисто механическим» математическим процессом, — т. е. любой *алгоритмический* процесс — может быть реализован в рамках конкретной схемы, открытой самим Черчем и названной им *лямбда-исчислением* ($\lambda$-исчислением)[11] (весьма, надо отметить, изящная и концептуально сдержанная схема; краткое ознакомительное изложение см. в НРК, с. 66–70). Вскоре после этого, в 1936–1937 годах, британский математик Алан Тьюринг нашел свой собственный, гораздо более убедительный способ описания алгоритмических процессов, основанный на функционировании теоретических «вычислительных машин», которые мы сейчас называем *машинами Тьюринга*. Вслед за Тьюрингом в некоторой степени аналогичную схему разработал американский ученый-логик польского происхождения Эмиль Пост (1936). Далее Черч и Тьюринг независимо друг от друга показали, что исчисление Черча эквивалентно концепции машины Тьюринга (а следовательно, и схеме Поста). Более того, именно этим концепциям Тьюринга в значительной степени обязаны своим появлением на свет современные универсальные компьютеры. Как уже упоминалось, машина Тьюринга по принципу функционирования фактически полностью эквивалентна современному компьютеру, — несколько, впрочем, идеализированному, т. е. обладающему возможностью использовать неограниченный объем памяти. Таким образом получается, что тезис Черча в его первоначальной формулировке всего лишь утверждает, что математическими алгоритмами следует считать как раз те процессы, которые способен выпол-

нить идеализированный современный компьютер — а если учесть общепринятое ныне *определение* термина «алгоритм», то такое утверждение и вовсе становится тавтологией. Так что принятие этой формулировки тезиса Черча не влечет за собой никакого противоречия точке зрения $\mathscr{C}$[4].

Вполне вероятно, однако, что сам Тьюринг имел в виду нечто большее: вычислительные возможности любого *физического* устройства должны (в идеале) быть эквивалентны действию машины Тьюринга. Такое утверждение существенно выходит за рамки того, что изначально подразумевал Черч. При разработке концепции «машины Тьюринга» сам Тьюринг основывался на своих представлениях о том, чего, в принципе, мог бы достичь вычислитель-человек (см. [198]). Судя по всему, он полагал, что физическое действие в общем (а под эту категорию подпадает и активность мозга человека) всегда можно свести к какой-либо разновидности действия машины Тьюринга. Быть может, это утверждение (физическое) следует называть «тезисом Тьюринга» — для того чтобы отличать его от оригинального «тезиса Черча», утверждения чисто математического, которому никоим образом не противоречит $\mathscr{C}$. Именно такой терминологии я намерен придерживаться далее в этой книге. Соответственно, точка зрения $\mathscr{C}$ противоречит в этом случае *тезису Тьюринга*, а вовсе не тезису Черча.

## 1.7. Хаос

В последние годы ученые проявляют огромный интерес к математическому феномену, известному под названием «хаос», — феномену, в рамках которого физические системы оказываются способными на якобы аномальное и непредсказуемое поведение (рис. 1.1). Образует ли феномен хаоса необходимую невычислимую физическую основу для такой точки зрения, как $\mathscr{C}$?

[4]Время от времени математики натыкаются на процедуру, которая «очевидно» алгоритмична по своей природе, пусть даже порой не всегда бывает ясно, как эту процедуру можно сформулировать в виде операций машины Тьюринга или лямбда-исчисления. В таких случаях можно утверждать, что, «согласно тезису Черча», такая операция и в самом деле должна существовать. См., например, [67]. В этом пути нет ничего зазорного, и, уж конечно, не возникает никакого противоречия с $\mathscr{C}$. Более того, на таком толковании тезиса Черча основывается большая часть рассуждений главы 3.



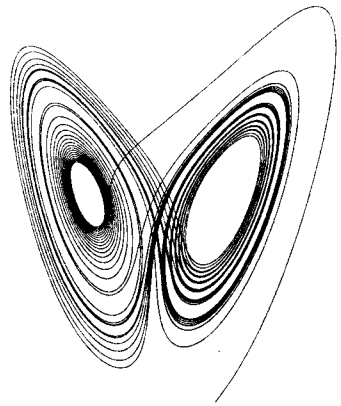


Рис. 1.1. Аттрактор Лоренца — один из первых примеров хаотической системы. Следуя линиям, мы переходим от левого лепестка аттрактора к правому и обратно произвольным, на первый взгляд, образом; то, в каком именно лепестке мы оказываемся в тот или иной момент времени, существенно зависит от нашей исходной точки. При этом кривая описывается простым математическим (дифференциальным) уравнением.

*Хаотические системы* — это динамически развивающиеся физические системы, математические модели таких физических систем или же просто математические модели, не описывающие никакой реальной физической системы и интересные сами по себе; характерно то, что будущее поведение такой системы чрезвычайно сильно зависит от ее начального состояния, причем определяющими могут оказаться самые незначительные факторы. Хотя обыкновенные хаотические системы являются полностью детерминированными и вычислительными, *на деле* может показаться, что в их поведении ничего детерминированного нет и никогда не было. Это происходит потому, что для сколько-нибудь надежного детерминистического предсказания будущего поведения системы необходимо знать ее начальное состояние с такой точностью, которая может оказаться просто недостижимой не только для тех измерительных средств, которыми мы располагаем, но также и для тех, которые мы только можем вообразить.

В этой связи чаще всего вспоминают о подробных долгосрочных прогнозах погоды. Законы, управляющие движением молекул воздуха, а также другими физическими величинами, которые могут оказаться релевантными для определения будущей погоды, хорошо известны. Однако реальные синоптические ситуации, которые могут возникнуть всего через несколько дней после предсказания, настолько тонко зависят от начальных условий, что нет никакой возможности измерить эти условия достаточно точно для того, чтобы дать хоть сколько-нибудь надежный прогноз. Безусловно, количество параметров, которые необходимо ввести в подобное вычисление, огромно; поэтому, быть может, и нет ничего удивительного в том, что в данном случае предсказание может оказаться на практике просто невозможным.

С другой стороны, подобное — так называемое хаотическое — поведение может иметь место и в случае очень простых систем; примером тому служат системы, состоящие из малого количества частиц. Вообразите, что от вас требуется загнать в лузу бильярдный шар E, расположенный пятым в некоторой извилистой[5] и очень растянутой цепочке шаров A, B, C, D и E; вам нужно ударить кием по шару A так, чтобы тот ударил шар B, который, в свою очередь, ударил бы шар C, который ударил бы шар D, который ударил бы шар E, который, наконец, попал бы в лузу. В общем случае необходимая для этого точность значительно превышает способности любого профессионального игрока в бильярд. Если бы цепочка состояла из 20 шаров, то тогда — даже допустив, что эти шары представляют собой идеально упругие точные сферы, — задача загнать в лузу последний шар оказалась бы не под силу и самому точному механизму из всех доступных современной технологии. Поведение последних шаров цепочки было бы, в сущности, случайным, несмотря на то, что управляющие поведением шаров ньютоновы законы математически абсолютно детерминированы и, в принципе, эффективно вычислимы. Никакое вычисление не смогло бы предсказать *реальное* поведение последних шаров цепочки просто потому, что нет никакой возможности добиться достаточно точного определения реального начального положения и скорости движения кия или положений первых шаров цепочки. Более того, даже самые незначительные внешние воздействия, вроде дыхания человека в соседнем городе, могут нарушить эту точность до такой степени, которая полностью обесценит результаты любого подобного вычисления.

[5] В черновом варианте книги слова «извилистой» здесь не было. Если шары расположены точно на прямой линии, этот трюк оказывается достаточно простым: я узнал об этом, к своему удивлению, когда попробовал проделать это сам. При расстановке шаров по прямой возникает неожиданная устойчивость, отсутствующая в общем случае.



Здесь необходимо пояснить, что, несмотря на столь серьезные трудности, встающие перед детерминистическим предсказанием, все нормальные системы, к которым применим термин «хаотические», *следует* относить к категории систем, которые я называю «вычислительными». Почему? Как и в других ситуациях, которые мы рассмотрим позднее, для того, чтобы определить, является ли та или иная процедура вычислительной, достаточно задать себе вопрос: выполнима ли она на обычном универсальном компьютере? Очевидно, что в данном случае ответ может быть только утвердительным, по той простой причине, что математически описываемые хаотические системы и в самом деле изучаются, как правило, с помощью компьютера!

Разумеется, если мы попытаемся создать компьютерную модель для подробного предсказания погоды в Европе в течение недели или же для описания последовательных столкновений расположенных вдоль некоторой кривой на достаточно большом расстоянии друг от друга двадцати бильярдных шаров после того, как по первому из них резко ударили кием, то можно почти с полной определенностью утверждать, что результаты, полученные с помощью нашей модели, и близко не будут похожи на то, что произойдет *в действительности*. Такова природа хаотических систем. На практике бесполезно пытаться с помощью вычислений предсказать *реальное* конечное состояние системы. Тем не менее, моделирование *типичного* конечного состояния вполне возможно. Предсказанная погода может и не совпасть с реальной, но она абсолютно правдоподобна как *погода вообще*! Точно так же и предсказанный результат столкновений бильярдных шаров абсолютно приемлем как *возможный* исход, даже несмотря на то, что на самом деле шары могут повести себя совершенно не так, как предсказано вычислением, — однако и при этом их поведение остается в равной степени приемлемым. Упомянем еще об одном обстоятельстве, которое подчеркивает идеально вычислительную природу таких операций: если запустить процесс компьютерного моделирования вторично, задав те же входные

данные, что и ранее, то результат моделирования будет *точно* таким же, как и в первый раз! (Здесь предполагается, что сам компьютер не ошибается; впрочем, надо признать, что современные компьютеры и в самом деле крайне редко совершают при вычислениях реальные ошибки.)

Возвращаясь к искусственному интеллекту, отметим, что никто пока и не пытается воспроизвести поведение какого-то конкретного индивидуума; нас бы прекрасно устроила модель *индивидуума вообще*! В этом контексте моя позиция вовсе не представляется такой уж неразумной: хаотические системы следует безусловно относить к категории систем, которые мы называем «вычислительными». Компьютерная модель такой системы и в самом деле выглядела бы как абсолютно приемлемый «типичный случай», даже и не совпадая при этом ни с каким «реальным случаем». Если внешние проявления человеческого разума суть результаты некоей хаотической динамической эволюции (эволюции вычислительной в том смысле, о котором мы только что говорили), то это вполне согласуется с точками зрения $\mathscr{A}$ и $\mathscr{B}$, но *никак не* $\mathscr{C}$.

Время от времени выдвигаются предположения, что, возможно, именно феномен хаоса — если, конечно, он действительно имеет место в деятельности мозга как физической сущности — позволяет человеческому мозгу *симулировать* поведение, якобы отличное от вычислительно-детерминированного функционирования машины Тьюринга, хотя, как подчеркивалось выше, формально его активность *является* целиком и полностью вычислительной. К этому вопросу мне еще придется вернуться несколько позднее (см. § 3.22). Пока же достаточно уяснить лишь то, что хаотические системы *относятся* к категории систем, называемых мною «вычислительными» или «алгоритмическими». Вопрос же о том, можно ли смоделировать какую-нибудь из таких систем *на практике*, не входит в круг *принципиальных* вопросов, которые мы здесь рассматриваем.

## 1.8. Аналоговые вычисления

До сих пор я рассматривал «вычисление» только в том смысле, в котором этот термин применим к современным цифровым компьютерам или, точнее, к их теоретическим предшественникам — машинам Тьюринга. Существуют и другие разновидности вычислительных устройств, особенно широко распространенные в не столь отдаленном прошлом; вычислительные операции здесь осуществляются не посредством переходов между дискретными состояниями «вкл./выкл.», знакомыми нам по цифровым вычислениям, а с помощью непрерывного изменения того или иного физического параметра. Самым известным из таких устройств является логарифмическая линейка, изменяемым физическим параметром которой является линейное расстояние (между фиксированными точками на линейке). Это расстояние служит для представления логарифмов чисел, которые нужно перемножить или разделить. Существует много различных разновидностей аналоговых вычислительных устройств, в которых могут применяться и другие типы физических параметров — такие, например, как время, масса или электрический потенциал.



В случае аналоговых систем необходимо учитывать одно формальное обстоятельство: стандартные понятия вычисления и вычислимости применимы, строго говоря, только к *дискретным* системам (над которыми, собственно, и выполняются «цифровые» действия), но не к *непрерывным*, таким, например, как расстояния или электрические потенциалы, с которыми имеет дело традиционная классическая физика. Иными словами, для того чтобы применить обычные вычислительные понятия к системе, описание которой требует не дискретных (или «цифровых»), а непрерывных параметров, мы естественным образом должны прибегнуть к аппроксимации. Действительно, при компьютерном моделировании физических систем вообще стандартной процедурой является *аппроксимация* всех рассматриваемых непрерывных параметров в дискретной форме. Подобная процедура, однако, неминуемо вносит некоторую погрешность, величина которой определяется заданной степенью точности аппроксимации; при этом вполне возможно, что для той или иной интересующей нас физической системы заданной точности может оказаться недостаточно. В итоге дискретное компьютерное моделирование очень просто может привести нас к ошибочным выводам относительно поведения моделируемой непрерывной физической системы.

В принципе, ничто не мешает повысить точность до уровня, адекватного для моделирования рассматриваемой непрерывной системы. Однако на практике, особенно в случае хаотических систем, требуемые для этого время вычислений и объем памяти могут оказаться непомерно большими. Кроме того, можем ли мы,

строго говоря, быть абсолютно уверены в том, что выбранная нами степень точности является *действительно* достаточной? Необходим какой-то критерий, который позволил бы нам определить, что нужный уровень точности достигнут, дальнейшего ее повышения не требуется и качественному поведению, вычисленному с такой точностью, в самом деле можно доверять. Все это поднимает ряд достаточно щекотливых математических вопросов, рассматривать которые подробно на этих страницах мне представляется не совсем уместным.

Существуют, однако, и другие подходы к проблемам вычислений в случае непрерывных систем; например, такие, в которых непрерывные системы рассматриваются как самостоятельные математические структуры со своим *собственным* понятием «вычислимости» — понятием, обобщающим идею вычислимости по Тьюрингу с дискретных величин на непрерывные[12]. При таком подходе исчезает необходимость в аппроксимации непрерывной системы дискретными параметрами с целью применить к ней традиционную концепцию вычислимости по Тьюрингу. Такие идеи вызывают определенный интерес с математической точки зрения; к сожалению, им, как нам представляется, не достает пока той неотразимой естественности и уникальности, которые присущи стандартному понятию вычислимости по Тьюрингу для дискретных систем. Более того, вследствие определенной непоследовательности данного подхода, формально «невычислимыми» оказываются и некоторые простые системы, в применении к которым подобная терминология выглядит как-то не совсем уместно (даже такие, например, как известное всем из физики простое «волновое уравнение»; см. [314] и НРК, с. 187–188). С другой стороны, следует упомянуть и об одной сравнительно недавней работе ([328]), в которой показано, что теоретические аналоговые компьютеры, объединяемые в некоторый достаточно обширный класс, не могут выйти за рамки обычной вычислимости по Тьюрингу. Я надеюсь, что дальнейшие исследования должным образом осветят эти безусловно интересные и важные темы. Пока же у меня нет оснований полагать, что работы в этом направлении в целом уже достигли той стадии завершенности, чтобы их результаты можно было применить к рассматриваемым здесь проблемам.

В этой книге меня в особенности занимает вопрос о вычислительной природе умственной деятельности, где термин «вычислительный» следует рассматривать в стандартном смысле *вычислимости по Тьюрингу*. В самом деле, компьютеры, которыми мы сегодня повседневно пользуемся, являются цифровыми, и именно это их свойство оказывается существенным для современных разработок в области ИИ. Наверное, логичным будет предположить, что в будущем может появиться «компьютер» какого-то иного типа, *решающую* роль в функционировании которого будут играть (пусть даже и не выходя при этом за общепринятые теоретические рамки современной физики) непрерывные физические параметры, что позволит такому компьютеру демонстрировать поведение, существенно *отличное* от поведения цифрового компьютера.



Как бы то ни было, все эти вопросы важны, главным образом, для проведения границы между «сильной» и «слабой» версиями позиции $\mathscr{C}$. Согласно *слабой* версии $\mathscr{C}$, поведение обладающего сознанием человеческого мозга обусловлено некоторой физической активностью, которую невозможно вычислить в стандартном смысле дискретной вычислимости по Тьюрингу, но которую можно полностью объяснить в рамках современных физических теорий. Если так, то эта активность, по всей видимости, должна зависеть от каких-то непрерывных физических параметров таким образом, чтобы ее невозможно было адекватно воспроизвести с помощью стандартных цифровых процедур. В соответствии же с *сильной* версией $\mathscr{C}$, невычислимость сознательной деятельности мозга может быть исчерпывающе объяснена в рамках некоторой невычислительной физической теории (пока еще не открытой), следствия из которой, собственно, и обусловливают упомянутую деятельность. Хотя второй вариант может показаться несколько надуманным, альтернатива (для сторонников $\mathscr{C}$) и в самом деле состоит в отыскании для какого-либо непрерывного процесса в рамках известных физических законов такой роли, которую невозможно было бы адекватно воспроизвести посредством каких угодно вычислений. На данный же момент, несомненно, следует ожидать, что для любой достоверной аналоговой системы любого типа из тех, что получили более или менее серьезное рассмотрение, *обязательно* окажется возможным (по крайней мере, в принципе) создать эффективную цифровую модель.

Даже если не принимать во внимание всевозможные теоретические проблемы общего плана, на сегодняшний день наиболь-

шее превосходство перед аналоговыми вычислительными системами демонстрируют именно *цифровые* компьютеры. Цифровые вычисления имеют гораздо более высокую точность благодаря, в основном, тому, что при хранении данных в цифровом виде повышение точности обеспечивается простым увеличением разрядности чисел, что легко достижимо с помощью весьма скромного увеличения (логарифмического) мощности компьютера; в аналоговых же машинах (по крайней мере, в *полностью* аналоговых, в конструкцию которых не заложено никаких цифровых концепций) увеличения точности можно добиться лишь посредством весьма и весьма значительного увеличения (линейного) соответствующих параметров. Возможно, когда-нибудь в будущем возникнут новые идеи, которые пойдут на пользу аналоговым вычислителям, однако в рамках современной технологии бóльшая часть существенных практических преимуществ принадлежит, по всей видимости, *цифровому* вычислению.

## 1.9. Невычислительные процессы

Из всех типов вполне определенных процессов, что приходят в голову, большая часть относится, соответственно, к категории феноменов, называемых мною «вычислительными» (имеются в виду, конечно же, «цифровые вычисления»). Возможно, читатель уже начал волноваться, что сторонники позиции $\mathscr{C}$ так и останутся у нас не при деле. Причем я еще ни словом не упоминал о строго *случайных* процессах, которые могут быть обусловлены, скажем, какими-либо исходными данными, получаемыми от квантовой системы. (О квантовой механике мы немного подробнее поговорим во второй части, главы 5 и 6.) Впрочем, для самой системы практически безразлично, подается на ее вход *подлинно* случайная последовательность данных или же всего лишь *псевдослучайная*, которую *можно* целиком и полностью сгенерировать вычислительным путем (см. §3.11). Действительно, несмотря на то, что между «случайным» и «псевдослучайным», строго говоря, существуют некоторые формальные отличия, они, на первый взгляд, не имеют непосредственного отношения к проблемам ИИ. Далее, в §3.11, §3.18 и последующих, я приведу некоторые серьезные доводы в пользу того, что «чистая случайность» и в самом деле абсолютно бесполезна для наших целей; если уж возникает такая необходимость, то лучше все же придерживаться псевдослучайности хаотического поведения, а все нормальные типы хаотического поведения, как уже подчеркивалось выше, относятся к категории «вычислительных».



А что нам известно о роли окружения? По мере развития каждого индивидуума у него или у нее формируется уникальное окружение, отличное от окружения любого другого человека. Возможно, именно это уникальное личное окружение и дает каждому из нас ту особенную последовательность входных данных, которая неподвластна вычислению? Хотя лично мне, например, сложно сообразить, на что именно в данном контексте может повлиять «уникальность» нашего окружения. Эти рассуждения напоминают разговор о хаосе, который мы вели выше (см. §1.7). Для обучения управляемого компьютером робота достаточно одной лишь модели некоего *правдоподобного* окружения (хаотического), при том, разумеется, условии, что в этой модели не будет ничего заведомо невычислимого. Роботу нет нужды учиться тем или иным навыкам в каком-то конкретном реальном окружении; его, разумеется, вполне устроит *типичное* окружение, моделирующее реальность вычислительными методами.

А может быть, численное моделирование пусть даже всего лишь правдоподобного окружения невозможно в принципе. Быть может, в окружающем физическом мире все же есть нечто такое, что на самом деле неподвластно численному моделированию. Возможно, некоторые сторонники $\mathscr{A}$ или $\mathscr{B}$ уже вознамерились приписать все не поддающиеся, на первый взгляд, вычислению проявления человеческого поведения невычислимости внешнего окружения. Должен, однако, заметить, что намерение это несколько опрометчиво. Ибо, как только мы признáем, что физическое поведение допускает *где-то* что-то такое, что невозможно моделировать вычислительными методами, мы тем самым тут же лишаемся главного, по всей видимости, основания сомневаться в правдоподобии, в первую очередь, самой точки зрения $\mathscr{C}$. Если во внешнем окружении (т. е. вне мозга) имеют место процессы, не поддающиеся численному моделированию, то почему не могут оказаться таковыми и процессы, протекающие *внутри* мозга? В конце концов, внутренняя физическая организация мозга человека, по всей видимости, гораздо более сложна, чем бóльшая часть (и это еще слабо сказано) его окружения, за исключением, быть может, тех его участков, где это окружение само оказывается под сильным влиянием деятельности других

мозгов. Признание возможности *внешней* невычислимой физической активности лишает всякой силы главный аргумент против $\mathscr{C}$. (См. также § 3.9, § 3.10.)

Следует сделать еще одно замечание относительно «не поддающихся вычислению» процессов, возможность существования которых предполагает позиция $\mathscr{C}$. Под этим термином я имею в виду *отнюдь* не те процессы, которые всего-навсего невычислимы *практически*. Здесь, конечно же, уместно вспомнить и о том, что, хотя моделирование любого правдоподобного окружения, или же любое точное воспроизведение всех физических и химических процессов, протекающих в мозге, может быть, в принципе, вычислимым, на такое вычисление, скорее всего, понадобится столько времени или такой объем памяти, что вряд ли удастся выполнить его на любом реально существующем или даже вообразимом в ближайшем будущем компьютере. Вероятно, нереально даже написание соответствующей компьютерной программы, если учесть, какое огромное количество различных факторов придется принимать в расчет. Однако сколь бы существенными ни были все эти соображения (а мы еще вернемся к ним в § 2.6, **Q8** и § 3.5), они не имеют *никакого* отношения к тому, что называю «невычислимостью» я (и чего требует $\mathscr{C}$). Под «невычислимостью» я подразумеваю *принципиальную* невозможность вычисления в том смысле, который мы очень скоро обсудим. Вычисления, которые просто выходят за рамки существующих (или вообразимых) компьютеров или имеющихся в нашем распоряжении вычислительных методов, формально все равно остаются «вычислениями».

Читатель имеет полное право спросить: если ничего, что можно счесть «невычислимым», не обнаруживается ни в случайности, ни во влиянии окружения, ни в банальном несоответствии уровня сложности феномена нашим техническим возможностям, то что вообще я имею в виду, говоря «чего требует $\mathscr{C}$»? В общем случае, это некий вид математически точной активности, невычислимость которой можно *доказать*. Насколько нам на данный момент известно, при описании физического поведения в подобной математической активности необходимости не возникает. Тем не менее, логически она возможна. Более того, она представляет собой нечто большее, нежели *просто* логическую возможность. Согласно приводимой далее в книге аргументации, возможность активности подобного общего характера *прямо* подразумевается



физическими законами, несмотря на то, что ни с чем подобным в известной физике мы еще не встречались. Некоторые примеры такой математической активности замечательно просты, поэтому представляется вполне уместным проиллюстрировать с их помощью то, о чем я здесь говорю.

Начать мне придется с описания нескольких примеров классов хорошо структурированных математических задач, не имеющих общего численного решения (ниже я поясню, в каком именно смысле). Начав с любого из таких классов задач, можно построить «игрушечную» модель физической вселенной, активность которой (даже будучи полностью детерминированной) фактически не поддается численному моделированию.

Первый пример такого класса задач знаменит более остальных и известен под названием «десятая проблема Гильберта». Эта задача была предложена великим немецким математиком Давидом Гильбертом в 1900 году в составе этакого перечня нерешенных на тот момент математических проблем, которые по большей части определили дальнейшее развитие математики в начале (да и в конце) двадцатого века. Суть десятой проблемы Гильберта заключалась в отыскании вычислительной процедуры, на основании которой можно было бы определить, имеют ли уравнения, составляющие данную систему *диофантовых* уравнений, хотя бы одно общее решение.

Диофантовыми называются полиномиальные уравнения с каким угодно количеством переменных, все коэффициенты и все решения которых должны быть *целыми числами*. (Целые числа — это числа, не имеющие дробной части, например: $\ldots, -3, -2, -1, 0, 1, 2, 3, 4, \ldots$. Первым такие уравнения систематизировал и изучил греческий математик Диофант в третьем веке нашей эры.) Ниже приводится пример системы диофантовых уравнений:

$$6w+2x^2-y^3=0, \quad 5xy-z^2+6=0, \quad w^2-w+2x-y+z-4=0.$$

Вот еще один пример:

$$6w+2x^2-y^3=0, \quad 5xy-z^2+6=0, \quad w^2-w+2x-y+z-3=0.$$

Решением первой системы является, в частности, следующее:

$$w=1, \quad x=1, \quad y=2, \quad z=4,$$

тогда как вторая система вообще не имеет решения (судя по первому уравнению, число $y$ должно быть четным, судя по второму уравнению, число $z$ также должно быть четным, однако это противоречит третьему уравнению, причем при любом $w$, поскольку значение разности $w^2 - w$ — это всегда четное число, а число 3 нечетно). Задача, поставленная Гильбертом, заключалась в отыскании математической процедуры (или *алгоритма*), позволяющей определить, какие системы диофантовых уравнений имеют решения (наш первый пример), а какие нет (второй пример). Вспомним (см. § 1.5), что алгоритм — это всего лишь вычислительная процедура, действие некоторой машины Тьюринга. Таким образом, решением десятой проблемы Гильберта является некая вычислительная процедура, позволяющая определить, когда система диофантовых уравнений имеет решение.

Десятая проблема Гильберта имеет очень важное историческое значение, поскольку, сформулировав ее, Гильберт поднял вопрос, который ранее не поднимался. Каков точный математический *смысл* словосочетания «алгоритмическое решение для класса задач»? Если точно, то что это вообще такое — «алгоритм»? Именно этот вопрос привел в 1936 году Алана Тьюринга к его собственному определению понятия «алгоритм», основанному на изобретенных им машинах. Примерно в то же время другие математики (Черч, Клин, Гёдель, Пост и др.; см. [135]) предложили несколько иные процедуры. Как вскоре было показано, все эти процедуры оказались эквивалентными либо определению Тьюринга, либо определению Черча, хотя особый подход Тьюринга приобрел все же наибольшее влияние. (Только Тьюрингу пришла в голову идея специфической и всеобъемлющей алгоритмической машины, — названной *универсальной* машиной Тьюринга, — которая способна самостоятельно выполнить абсолютно *любое* алгоритмическое действие. Именно эта идея привела впоследствии к созданию концепции универсального компьютера, который сегодня так хорошо нам знаком.) Тьюрингу удалось показать, что существуют определенные классы задач, которые *не имеют* алгоритмического решения (в частности, «проблема остановки», о которой я расскажу ниже). Однако самóй десятой проблеме Гильберта пришлось ждать своего решения до 1970 года, когда русский математик Юрий Матиясевич (представив доказательства, ставшие логическим завершением некоторых соображений, выдвинутых ранее американскими математиками Джулией Робинсон, Мартином Дэвисом и Хилари Патнэмом) показал невозможность создания компьютерной программы (или алгоритма), способной систематически определять, имеет ли решение та или иная система диофантовых уравнений. (См. [72] и [89], глава 6, где приводится весьма занимательное изложение этой истории.) Заметим, что в случае утвердительного ответа (т. е. когда система имеет-таки решение), этот факт, в принципе, можно констатировать с помощью особой компьютерной программы, которая самым тривиальным образом проверяет один за другим все возможные наборы целых чисел. Сколько-нибудь систематической обработке не поддается именно случай отсутствия решения. Можно, конечно, создать различные совокупности правил, которые корректно определяли бы, когда система не имеет решения (наподобие приведенного выше рассуждения с использованием четных и нечетных чисел, исключающего возможность решения второй системы), однако, как показывает теорема Матиясевича, список таких совокупностей *никогда* не будет полным.



Еще одним примером класса вполне структурированных математических задач, не имеющих алгоритмического решения, является *задача о замощении*. Она формулируется следующим образом: дан набор многоугольников, требуется определить, покрывают ли они плоскость; иными словами, возможно ли покрыть всю евклидову плоскость только этими многоугольниками без зазоров и наложений? В 1966 году американский математик Роберт Бергер показал (причем эффективно), что эта задача вычислительными средствами неразрешима. В основу его доводов легло обобщение одной из работ американского математика китайского происхождения Хао Вана, опубликованной в 1961 году (см. [176]). Надо сказать, что в моей формулировке задача оказывается несколько более громоздкой, чем хотелось бы, так как многоугольные плитки описываются в общем случае с помощью вещественных чисел (чисел, выражаемых в виде бесконечных десятичных дробей), тогда как обычные алгоритмы способны оперировать только целыми числами. От этого неудобства можно избавиться, если в качестве рассматриваемых многоугольников выбрать плитки, состоящие из нескольких квадратов, примыкающих один к другому сторонами. Такие плитки называются *полиомино* (см. [161]; [136], глава 13; [222]). На рис. 1.2 показаны некоторые плитки полиомино и примеры замощений ими плоскости.

(Другие примеры замощений плоскости наборами плиток см. в НРК, с. 133–137, рис. 4.6–4.12.) Любопытно, что вычислительная неразрешимость задачи о замощении связана с существованием наборов полиомино, называемых *апериодическими*; такие наборы покрывают плоскость *исключительно апериодически* (т. е. так, что никакой участок законченного узора нигде не повторяется, независимо от площади покрытой плиткой плоскости). На рис. 1.3 представлен апериодический набор из трех полиомино (полученный из набора, обнаруженного Робертом Амманом в 1977 году; см. [176], рис. 10.4.11–10.4.13 на с. 555–556).

Математические доказательства неразрешимости с помощью вычислительных методов десятой проблемы Гильберта и задачи о замощении весьма сложны, и я, разумеется, не стану и пытаться приводить их здесь[13]. Центральное место в каждом из этих доказательств отводится, в сущности, тому, чтобы показать, каким образом можно запрограммировать машину Тьюринга на решение задачи о диофантовых уравнениях или задачи о замощении. В результате все сводится к вопросу, который Тьюринг рассматривал еще в своем первоначальном исследовании: к вычислительной неразрешимости *проблемы остановки* — проблемы определения ситуаций, в которых работа машины Тьюринга не может завершиться. В §2.3 мы приведем несколько примеров явных вычислительных процедур, которые принципиально *не могут* завершиться, а в §2.5 будет представлено достаточно простое доказательство — основанное, по большей части, на оригинальном доказательстве Тьюринга, — которое, помимо прочего, показывает, что проблема остановки действительно неразрешима вычислительными методами. (Что же касается следствий из того самого «прочего», ради которого, собственно, и затевалось упомянутое доказательство, то на них, в сущности, построены рассуждения всей первой части книги.)

Каким же образом можно применить такой класс задач, как задачи о диофантовых уравнениях или задачи о замощении, к созданию «игрушечной» вселенной, которая, будучи детерминированной, является, тем не менее, невычислимой? Допустим, что в нашей модели вселенной течет *дискретное время*, параметризованное натуральными (т. е. целыми неотрицательными) числами 0, 1, 2, 3, 4, .... Предположим, что в некий момент времени $n$ состояние вселенной точно определяется одной задачей из рассматриваемого класса, скажем, набором полиомино. Необходи-



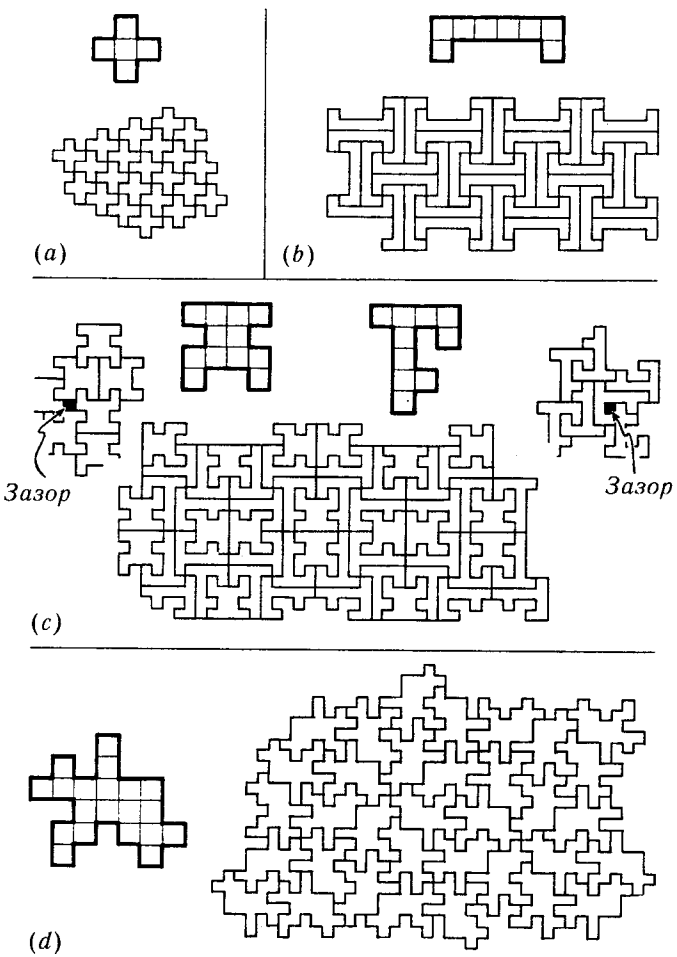



Рис. 1.2. Плитки полиомино и замощения ими бесконечной евклидовой плоскости (допускается использование зеркально отраженных плиток). Если брать полиомино из набора (c) по отдельности, то ни одно из них не покроет всю плоскость.

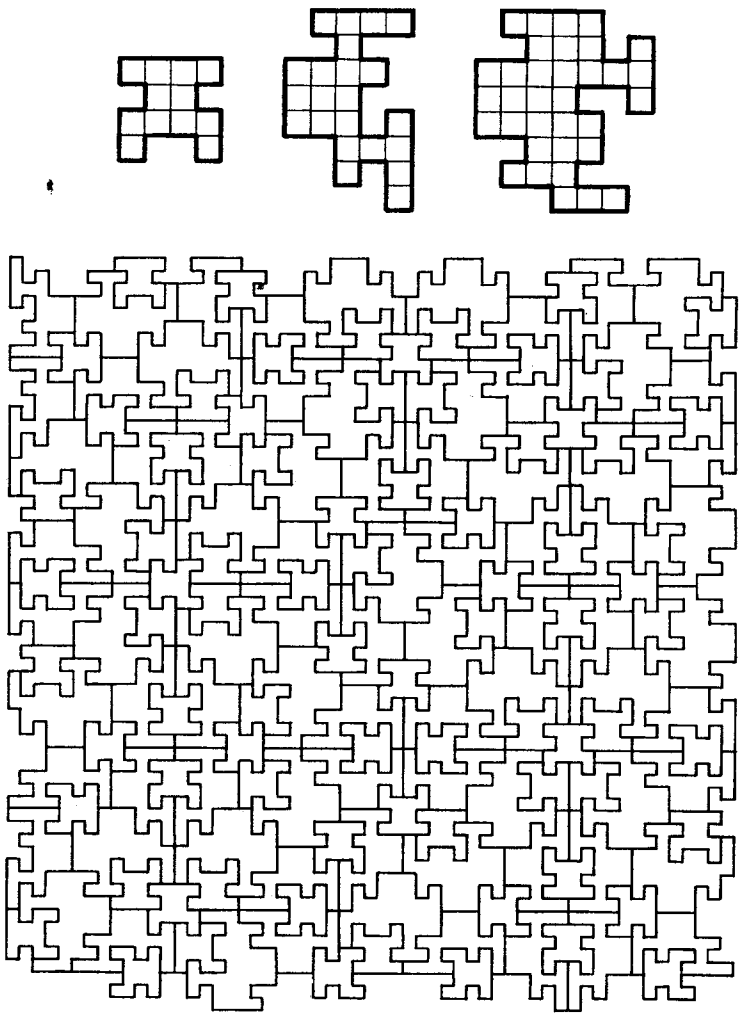


Рис. 1.3. Набор из трех полиомино, покрывающий плоскость апериодически (получен из набора Роберта Аммана).



мо установить два вполне определенных правила относительно того, какой из наборов полиомино будет представлять состояние вселенной в момент времени $n + 1$ при заданном наборе полиомино для состояния вселенной в момент времени $n$, причем первое из этих правил применяется в том случае, если полиомино *покрывают* всю плоскость без зазоров и наложений, а второе — если это *не так*. То, как именно будут выглядеть подобные правила, не имеет в данном случае особого значения. Можно составить список $S_0$, $S_1$, $S_2$, $S_3$, $S_4$, $S_5$, ... всех возможных наборов полиомино таким образом, чтобы наборы, содержащие в общей сложности *четное* число квадратов, имели бы четные индексы $S_0$, $S_2$, $S_4$, $S_6$, ..., а наборы с *нечетным* количеством квадратов — нечетные индексы $S_1$, $S_3$, $S_5$, $S_7$, .... (Составление такого списка не представляет особой сложности; нужно лишь подобрать соответствующую вычислительную процедуру.) Итак, «динамическая эволюция» нашей игрушечной вселенной задается теперь следующим условием:

> Из состояния $S_n$ в момент времени $t$ вселенная переходит в момент времени $t + 1$ в состояние $S_{n+1}$, если набор полиомино $S_n$ *покрывает* плоскость, и в состояние $S_{n+2}$, если набор $S_n$ *не покрывает* плоскость.

Поведение такой вселенной полностью детерминировано, однако поскольку в нашем распоряжении нет общей вычислительной процедуры, позволяющей установить, какой из наборов полиомино $S_n$ покрывает плоскость (причем это верно и тогда, когда общее число квадратов постоянно, независимо от того, четное оно или нет), то невозможно и численное моделирование ее реального развития. (См. рис. 1.4.)

Безусловно, такую схему нельзя воспринимать хоть сколько-нибудь всерьез — она ни в коем случае не моделирует реальную вселенную, в которой все мы живем. Эта схема приводится здесь (как, собственно, и в НРК, с. 170) для иллюстрации того часто недооцениваемого факта, что между детерминизмом и вычислимостью существует вполне определенная разница. *Некоторые полностью детерминированные модели вселенной с четкими законами эволюции невозможно реализовать вычислительными средствами.* Вообще говоря, как мы убедимся в §7.9, только что рассмотренные мною весьма специфические модели не совсем отвечают реальным требованиям точки зре-

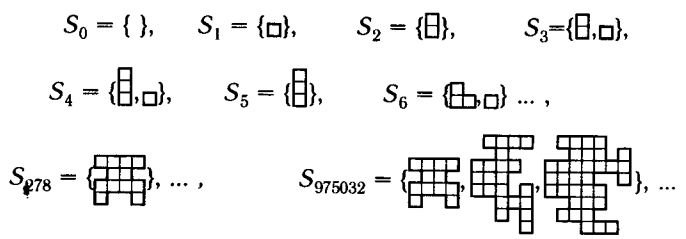



Рис. 1.4. Невычислимая модель «игрушечной» вселенной. Различные состояния этой детерминированной, но невычислимой вселенной даны в виде возможных конечных наборов полиомино, пронумерованных таким образом, что четные индексы $S_n$ соответствуют четному общему количеству квадратов в наборе, а нечетные индексы — нечетному количеству квадратов. Временна́я эволюция происходит в порядке увеличения индекса ($S_0$, $S_2$, $S_3$, $S_4$, ..., $S_{278}$, $S_{280}$, ...), при этом индекс пропускается, когда предыдущий набор оказывается не в состоянии замостить плоскость.

ния $\mathscr{C}$. Что же касается тех феноменов, которые отвечают-таки этим самым *реальным* требованиям, и некоторых связанных с упомянутыми феноменами поразительных физических возможностях, то о них мы поговорим в § 7.10.

## 1.10. Завтрашний день

Так какого же будущего для этой планеты нам следует ожидать согласно точкам зрения $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$, $\mathscr{D}$? Если верить $\mathscr{A}$, то настанет время, когда соответствующим образом запрограммированные суперкомпьютеры догонят — а затем и перегонят — человека во всех его интеллектуальных достижениях. Конечно же, сторонники $\mathscr{A}$ придерживаются различных взглядов относительно необходимого для этого времени. Некоторые вполне разумно полагают, что пройдет еще много столетий, прежде чем компьютеры достигнут уровня человека, принимая во внимание крайнюю скудость современного понимания реально выполняемых мозгом вычислений (так они говорят), обусловливающих ту тонкость поведения, какую, несомненно, демонстрирует человек, — тонкость, без которой, конечно же, нельзя говорить о каком бы то ни было «пробуждении сознания». Другие утверждают, что времени понадобится значительно меньше. В частности, Ханс Моравек в своей книге «Дети разума» [267] приводит вполне аргументированное доказательство (основанное на непрерывно ускоряющемся развитии компьютерных технологий за последние пятьдесят лет и на своей оценке той доли от всего объема функциональной активности мозга, которая на сегодняшний день уже успешно моделируется численными методами) в поддержку своего утверждения, будто уровень «эквивалентности человеку» будет преодолен уже к 2030 году. (Кое-кто утверждает, что это время будет еще короче[14], а кто-то даже уверен, что предсказанная дата достижения эквивалентности человеку уже осталась в прошлом!) Однако чтобы читатель не очень пугался того, что менее чем через сорок (или около того) лет компьютеры во всем его превзойдут, горькая пилюля подслащена одной радужной надеждой (подаваемой под видом гарантированного обещания): все мы сможем тогда перенести свои «ментальные программы» в сверкающие металлические (или пластиковые) корпуса роботов (конкретную модель, разумеется, каждый выберет себе сам), чем и обеспечим себе что-то вроде бессмертия [267, 268].



А вот для сторонников точки зрения $\mathscr{B}$ подобный оптимизм — непозволительная роскошь. Они вполне согласны с приверженцами $\mathscr{A}$ относительно перспектив развития интеллектуальных способностей компьютеров — с той лишь оговоркой, что речь при этом идет исключительно о внешних проявлениях этих самых способностей. Для управления роботом необходимо и достаточно располагать адекватной *моделью* деятельности человеческого мозга, больше ничего не требуется (рис. 1.5). Согласно $\mathscr{B}$, вопрос о том, способно ли подобное моделирование вызвать осмысленное осознание, не имеет никакого отношения к реальному поведению робота. На достижение необходимого для такого моделирования технологического уровня может уйти как несколько веков, так и менее сорока лет. Однако, как уверяют сторонники $\mathscr{B}$, рано или поздно, а это все-таки произойдет. Тогда же компьютеры достигнут уровня «эквивалентности человеку», а затем, как можно ожидать, и уверенно превзойдут его, оставив без внимания все потуги нашего относительно слабого мозга хоть немного этот уровень приподнять. Причем возможности «подключения» к управляемым роботам у нас в этом случае не будет,

и, похоже, придется примириться с тем, что нашей планетой, в конечном итоге, будут править абсолютно бесчувственные машины! Мне представляется, что из всех точек зрения $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$ и $\mathscr{D}$ именно $\mathscr{B}$ предлагает самый пессимистичный взгляд на будущее нашей планеты — вопреки, казалось бы, тому факту, что именно она лучше всего соотносится с так называемым «здравым смыслом».

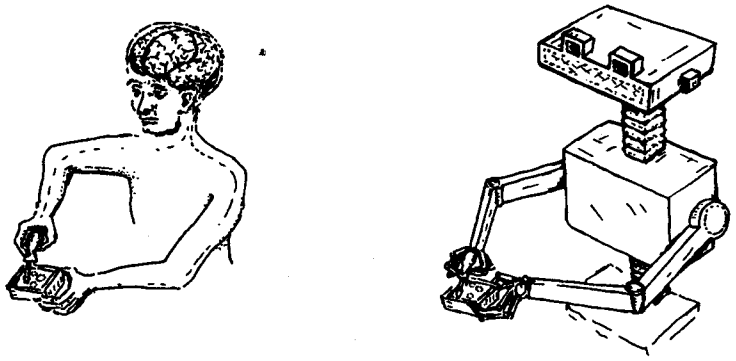


Рис. 1.5. Согласно точке зрения $\mathscr{B}$, компьютерное моделирование деятельности самосознающего человеческого мозга, в принципе, возможно; поэтому, в конечном итоге, управляемые компьютером роботы смогут догнать — а затем и значительно обогнать — человека во всех его интеллектуальных достижениях.

Если же верить $\mathscr{C}$ или $\mathscr{D}$, то можно ожидать, что компьютеры навсегда сохранят подчиненное по отношению к человеку положение — какими бы быстрыми, мощными или алгоритмически совершенными они ни стали. При этом точка зрения $\mathscr{C}$ не отрицает возможности будущих научных разработок, которые могут привести к созданию неких устройств, принцип действия которых *не* будет иметь ничего общего с компьютерами в их сегодняшнем понимании, а будет основан на той самой невычислимой физической активности, которая, согласно $\mathscr{C}$, обусловливает наше собственное сознательное мышление, — устройств, которые окажутся способны вместить в себя *реальные* разум и сознание. Быть может, в конечном итоге именно *такие* устройства, а вовсе не те машины, которые мы называем «компьютерами»,



и превзойдут человека в интеллектуальном отношении. Что ж, не исключено; однако подобные умозрительные прогнозы представляются мне в настоящий момент крайне преждевременными, поскольку мы практически не обладаем необходимыми для таких исследований научными познаниями, не говоря уже о каких бы то ни было технологических решениях. К этому вопросу мы еще вернемся во второй части книги (§ 8.1).

## 1.11. Обладают ли компьютеры правами и несут ли ответственность?

С некоторых пор умы теоретиков от *юриспруденции* начал занимать один вопрос, имеющий самое непосредственное отношение к теме нашего разговора, но в некотором смысле более практический[15]. Суть его заключается в следующем: не предстоит ли нам в не столь отдаленном будущем задуматься над тем, обладают ли компьютеры законными правами и несут ли они ответственность за свои действия. В самом деле, если со временем компьютеры смогут достичь уровня человека (а то и превзойти его) в самых разных областях деятельности, то подобные вопросы неминуемо должны приобрести определенную значимость. Если придерживаться точки зрения $\mathscr{A}$, то следует, очевидно, признать, что компьютеры (или управляемые компьютером роботы) должны потенциально и обладать правами, и нести ответственность. Ибо, согласно этой точке зрения, между человеком и роботом достаточно высокого уровня сложности нет существенной разницы, за исключением такой «мелочи», как различие в материальном строении. Однако приверженцам точки зрения $\mathscr{B}$ ситуация представляется несколько более запутанной. Разумно утверждать, что вопрос о правах или ответственности уместен для созданий, наделенных способностью чувствовать, т. е. испытывать определенные, подлинно душевные «ощущения» — такие, как страдание, гнев, мстительность, злоба, вера (религиозная и общечеловеческая), желание, сомнение, понимание или страсть. Согласно $\mathscr{B}$, управляемый компьютером робот не обладает такой способностью, вследствие чего, на мой взгляд, не может ни обладать правами, ни нести ответственность. С другой стороны, если верить $\mathscr{B}$, не существует эффективного способа определить, что упомянутая способность у робота действительно отсутствует, поэтому если роботы смогут достаточно правдоподобно имити-

ровать поведение человека, то человек может оказаться в весьма затруднительном положении.

Подобного затруднения, по всей видимости, не возникнет у сторонников точки зрения $\mathscr{C}$ (а также, возможно, $\mathscr{D}$), поскольку, согласно этим точкам зрения, компьютеры не в состоянии убедительно *демонстрировать* душевные переживания и, уж конечно же, ничего похожего не чувствуют и чувствовать никогда не будут. Соответственно, компьютеры *не могут ни* обладать правами, *ни* нести ответственность. Лично мне такая точка зрения представляется весьма разумной. Вообще в этой книге я выступаю как серьезный противник позиций $\mathscr{A}$ и $\mathscr{B}$. Согласившись с моими аргументами, юристы, безусловно, существенно упростят себе жизнь: как таковые компьютеры или управляемые компьютерами роботы *ни при каких обстоятельствах* не обладают правами и не несут ответственности. Нельзя обвинить компьютеры в каких бы то ни было неприятностях или недоразумениях — виновен всегда человек!

Следует, однако, понимать, что вышеприведенные аргументы могут и не относиться к всевозможным гипотетическим «устройствам», подобным упомянутым выше — тем, что смогут в конечном итоге воплотить в себе принципы новой, невычислительной физики. Но поскольку перспектива появления таких устройств — если их вообще удастся создать — весьма туманна, возникновения связанных с ними юридических проблем в ближайшем будущем ожидать не приходится.

Проблема «ответственности» поднимает глубокие философские вопросы, связанные с основными факторами, обусловливающими наше поведение. Можно вполне обоснованно утверждать, что каждое наше действие так или иначе определяется наследственностью и окружением, а то и всевозможными случайностями, непрерывно влияющими на нашу жизнь. Но ведь *ни одно* из этих воздействий никак не зависит лично от нас, почему же *мы* должны нести за них ответственность? Является ли понятие «ответственности» лишь терминологической условностью, или дело в чем-то еще? Возможно, и впрямь существует некая «самость» — нечто, стоящее «выше» уровня подобных влияний и определяющее, в конечном счете, наши действия? В юридическом смысле понятие «ответственности» *явно* подразумевает, что внутри каждого из нас и в самом деле существует своего рода независимая «самость», наделенная своей *собственной* ответственностью — и, по определению, правами, — причем ее проявления *нельзя* объяснить ни наследственностью, ни окружением, ни случайностью. Если же присутствие в нашей речи такой независимой «самости» не просто языковая условность, то в современных физических представлениях недостает чего-то весьма существенного. Открытие этого недостающего ингредиента, несомненно, многое изменит в нашем научном мировоззрении.



Хотя книга, которую вы держите в руках, и не дает исчерпывающего ответа на эти серьезные вопросы, она, как я полагаю, может чуть приоткрыть дверь, отделяющую нас от него, — не больше, но и не меньше. Вы не найдете здесь неопровержимых доказательств непременного существования такой «самости», проявления которой нельзя объяснить никакой внешней причиной, вам лишь предложат несколько шире взглянуть на саму природу возможных «причин». «Причина» может оказаться невычислимой — на практике или в принципе. Я намерен показать, что если упомянутая «причина» так или иначе порождается нашими сознательными действиями, то она должна быть весьма тонкой, безусловно невычислимой и не имеющей ничего общего ни с хаосом, ни с прочими чисто случайными воздействиями. Сможет ли такая концепция «причины» приблизить нас к пониманию истинной сущности свободы воли (или иллюзорности такой свободы) — вопрос будущего.

## 1.12. «Осознание», «понимание», «сознание», «интеллект»

До сих пор я не ставил перед собой задачи точно определить те неуловимые концепции, что так или иначе связаны с проблемой «разума». Формулируя положения $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$ и $\mathscr{D}$ в § 1.3, я несколько туманно упоминал об «осознании», других же свойств мышления мы пока не касались. Думаю, что следует хотя бы попытаться прояснить используемую здесь и далее терминологию — особенно в отношении таких понятий, как «понимание», «сознание» и «интеллект», играющих весьма существенную роль в наших рассуждениях.

Хотя я не вижу особой необходимости пытаться дать непременно полные определения, некоторые комментарии относитель-

но моей собственной терминологии представляются все же уместными. Я часто с некоторым замешательством обнаруживаю, что употребление всех этих слов, столь очевидное для меня, не совпадает с тем, что полагают естественным другие. Например, термин «понимание», на мой взгляд, безусловно подразумевает, что истинное обладание этим свойством требует некоторого элемента *осознания*. Не осознав сути того или иного суждения, мы, разумеется, не можем претендовать на истинное понимание этого самого суждения. По крайней мере, *я* уверен, что эти слова следует понимать именно так, хотя провозвестники ИИ, похоже, со мною не согласны и используют термины «понимание» и «осознание» в некоторых контекстах так, что первое никоим образом не предполагает непременного наличия второго. Некоторые из них (принадлежащие к категории $\mathscr{A}$ или $\mathscr{B}$) полагают, что управляемый компьютером робот «понимает», в чем заключаются его инструкции, однако при этом никто и не заикается о том, что робот свои инструкции действительно «осознает». Мне кажется, что здесь перед нами всего-навсего неверное употребление термина «понимание», пусть даже одно из тех, что обладают подлинной эвристической ценностью для описания функционирования компьютера. Когда мне потребуется указать на то, что термин «понимание» используется не в таком эвристическом смысле — т. е. при описании деятельности, для которой действительно необходимо осознание, — я буду использовать сочетание «подлинное понимание».

Кое-кто, разумеется, может заявить, что между этими двумя случаями употребления слова «понимание» нет четкого различия. Если это так, то сама концепция осознания также не имеет точного определения. С этим, конечно, не поспоришь; однако у меня нет никаких сомнений в том, что осознание действительно представляет собой некоторую *сущность*, причем эта сущность может как наличествовать, так и отсутствовать, — по крайней мере, до некоторой степени. Если согласиться с тем, что осознание представляет-таки собой некоторую сущность, то вполне естественно будет согласиться и с тем, что эта сущность должна являться неотъемлемой частью всякого подлинного понимания. Это утверждение, кстати, не отрицает возможности того, что «сущность», которой является осознание, окажется в действительности результатом чисто вычислительной деятельности в полном соответствии с точкой зрения $\mathscr{A}$.



Я также полагаю, что термин «интеллект» следует употреблять исключительно в связи с пониманием. Некоторые же теоретики от ИИ берутся утверждать, что их робот вполне может обладать «интеллектом», не испытывая при этом никакой необходимости в действительном «понимании» чего-либо. Термин «искусственный интеллект» предполагает возможность осуществления разумной вычислительной деятельности, и, вместе с тем, многие полагают, что разрабатываемый ими ИИ замечательно обойдется без подлинного понимания — и, как следствие, осознания. На мой взгляд, словосочетание «интеллект без понимания» есть лишь результат неверного употребления терминов. Следует, впрочем, отметить, что иногда что-то вроде частичного моделирования подлинного интеллекта без какого бы то ни было реального понимания оказывается до определенной степени возможным. (В самом деле, не так уж редко встречаются *человеческие* существа, способные на некоторое время одурачить нас демонстрацией какого-никакого понимания, хотя, как в конце концов выясняется, оно им в принципе не свойственно!) Между подлинным интеллектом (или подлинным пониманием) и любой деятельностью, моделируемой исключительно вычислительными методами, действительно существует четкое различие; это утверждение является одним из важнейших положений моих дальнейших рассуждений. Согласно моей терминологии, обладание *подлинным* интеллектом непременно предполагает присутствие подлинного понимания. То есть, употребляя термин «интеллект» (особенно в сочетании с прилагательным «подлинный»), я тем самым подразумеваю наличие некоторого действительного осознания.

Лично мне такая терминология кажется совершенно естественной, однако многие поборники ИИ (во всяком случае те из них, кто *не* поддерживает точку зрения $\mathscr{A}$) станут решительно отрицать всякую свою причастность к попыткам реализации искусственного «осознания», хотя конечной их целью является, судя по названию, не что иное, как искусственный «интеллект». Они, пожалуй, оправдаются тем, что они (в полном согласии с $\mathscr{B}$) всего лишь *моделируют* интеллект — такая модель не требует *действительного* понимания или осознания, — а вовсе не пытаются создать то, что я называю *подлинным* интеллектом. Вероятно, они будут уверять вас, что не видят никакой разницы между подлинным интеллектом и его моделью, что вполне отвеча-

ет точке зрения $\mathscr{A}$. В своих дальнейших рассуждениях я, в частности, намерен показать, что некоторые аспекты «подлинного понимания» действительно невозможно воссоздать путем каких бы то ни было вычислений. Следовательно, должно существовать и различие между подлинным интеллектом и любой попыткой его достоверного численного моделирования.

Я, разумеется, не даю определений ни «интеллекту», ни «пониманию», ни, наконец, «осознанию». Я полагаю в высшей степени неблагоразумным пытаться дать в рамках данной книги *полное* определение *хотя бы одному* из упомянутых понятий. Нам придется до некоторой степени положиться на свое интуитивное восприятие действительного смысла этих слов. Если интуиция подсказывает нам, что «понимание» есть нечто, необходимое для «интеллекта», то любое доказательство невычислительной природы «понимания» автоматически доказывает и невычислительную природу «интеллекта». Более того, если «пониманию» непременно должно предшествовать «осознание», то невычислительное физическое обоснование феномена осознания вполне в состоянии объяснить и аналогичную невычислительную природу «понимания». Итак, мое употребление этих терминов (в сущности совпадающее, как я полагаю, с общеупотребительным) сводится к двум положениям:

а) «интеллект» *требует* «понимания»

и

б) «понимание» *требует* «осознания».

Осознание я воспринимаю как один из аспектов — *пассивный* — феномена *сознания*. У сознания имеется и *активный* аспект, а именно — *свободная воля*. Полного определения слова «сознание» здесь также не дается (и, уж конечно же, не мне определять, что есть «свободная воля»), хотя мои аргументы имеют целью окончательное объяснение феномена сознания в научных, но невычислительных терминах — как того требует точка зрения $\mathscr{C}$. Не претендую я и на то, что мне удалось преодолеть хоть сколько-нибудь значительное расстояние на пути к этой цели, однако надеюсь, что представленная в этой книге (равно как и в НРК) аргументация расставит вдоль этого пути несколько полезных указателей для идущих следом — а может, станет и чем-то бóльшим. Мне кажется, что, пытаясь на данном



этапе дать слишком точное определение термину «сознание», мы рискуем упустить ту самую концепцию, какую хотим изловить. Поэтому вместо поспешного и наверняка неадекватного определения я приведу лишь несколько комментариев описательного характера относительно моего собственного употребления термина «сознание». В остальном же нам придется положиться на интуитивное понимание смысла этого термина.

Все это вовсе не означает, что я полагаю, будто мы действительно «интуитивно знаем», чем на самом деле «является» сознание; я лишь хочу сказать, что такое понятие существует, а мы, по мере сил, пытаемся его постичь — причем за понятием стоит некий реально существующий феномен, который допускает научное описание и играет в физическом мире как пассивную, так и активную роль. Некоторые, судя по всему, полагают, что данная концепция слишком туманна, чтобы заслуживать серьезного изучения. Однако при этом те же люди[16] часто и с удовольствием рассуждают о «разуме», полагая, очевидно, что это понятие определено гораздо точнее. Общепринятое употребление слова «разум» предполагает разделение этого самого разума (возможное или реальное) на так называемые «сознательную» и «бессознательную» составляющие. На мой взгляд, концепция бессознательного разума представляется еще более невразумительной, нежели концепция разума сознательного. Я и сам нередко пользуюсь словом «разум», однако не пытаюсь при этом дать его точное определение. В нашей последующей дискуссии (достаточно строгой, надеюсь) концепция «разума» — *за исключением* той ее части, что уже нашла свое воплощение в термине «сознание», — не будет играть центральной роли.

Что же я имею в виду, говоря о сознании? Как уже отмечалось ранее, сознание обладает активным и пассивным аспектами, однако различие между ними далеко не всегда четко определено. Восприятие, скажем, красного цвета требует несомненно пассивного сознания, равно как и ощущение боли либо восхищение музыкальным произведением. Активное же сознание участвует в сознательных действиях — таких, например, как подъем с кровати или, напротив, намеренное решение воздержаться от какой-либо энергичной деятельности. При воссоздании в памяти каких-то прошедших событий оказываются задействованы как пассивный, так и активный аспекты сознания. Составление плана будущих действий также обычно требует участия сознания —

и активного, и пассивного; и, надо полагать, какое-никакое сознание необходимо для умственной деятельности, которую общепринято описывать словом «понимание». Более того, мы остаемся, в определенном смысле, в сознании (пассивный аспект), даже когда спим, если при этом нам снится сон (в процессе же пробуждения может принимать участие и активный аспект сознания).

У кого-то могут найтись возражения против того, что все эти разнообразные проявления сознания следует загонять в тесные рамки какой-то одной — пусть и всеобъемлющей — концепции. Можно, например, указать на то, что для описания феномена сознания необходимо принимать во внимание множество самых разных концепций, не ограничиваясь простым разделением на «активное» и «пассивное», а также и то, что реально существует огромное количество различных психических признаков, каждый из которых имеет определенное отношение к тому или иному свойству мышления. Соответственно, применение ко всем этим свойствам общего термина «сознание» представляется, в лучшем случае, бесполезным. Мне все же думается, что должна существовать некая единая концепция «сознания», центральная для всех отдельных аспектов мыслительной деятельности. Говоря о разделении сознания на пассивный и активный аспекты (иногда четко отличимые один от другого, причем пассивный аспект связан с ощущениями (или *qualia*), а активный — с проявлениями «свободной воли»), я считаю их двумя сторонами одной монеты.

В первой части книги меня будет занимать, главным образом, вопрос о том, чего можно достичь, используя свойство мышления, известное как «понимание». Хотя я не даю здесь определения термину «понимание», надеюсь все же прояснить его смысл в достаточной мере для того, чтобы убедить читателя в том, что обозначаемое этим термином свойство — чем бы оно ни оказалось — и в самом деле должно быть неотъемлемой частью мыслительной деятельности, которая необходима, скажем, для признания справедливости рассуждений, составляющих § 2.5. Я намерен показать, что восприятие *этих* рассуждений должно быть связано с какими-то принципиально невычислимыми процессами. Мое доказательство не затрагивает столь *непосредственно* другие свойства мыслительной деятельности («интеллект», «осознание», «сознание» или «разум»), однако оно имеет определенное отношение и к этим концепциям, поскольку, в соответствии



с той терминологией «от здравого смысла», о которой я упоминал выше, осознание непременно должно быть существенным компонентом понимания, а понимание — являться неотъемлемой частью любого подлинного интеллекта.

## 1.13. Доказательство Джона Серла

Прежде чем представить свое собственное рассуждение, хотелось бы упомянуть о совсем иной линии доказательства — знаменитой «китайской комнате» философа Джона Серла[17] — главным образом для того, чтобы подчеркнуть существенное отличие от нее моего доказательства как по общему характеру, так и по базовым концепциям. Доказательство Серла тоже связано с проблемой «понимания» и имеет целью выяснить, можно ли утверждать, что функционирование достаточно сложного компьютера реализует это свойство мышления. Я не буду повторять здесь рассуждение Серла во всех подробностях, а лишь кратко обозначу его суть.

Дана некая компьютерная программа, которая демонстрирует имитацию «понимания», отвечая на вопросы о какой-то рассказанной ей предварительно истории, причем все вопросы и ответы даются на китайском языке. Далее Серл рассматривает не владеющего китайским языком человека, который старательно воспроизводит все до единой вычислительные операции, выполняемые в процессе имитации компьютером. Когда вычисления выполняет *компьютер*, получаемые на его выходе данные создают некоторую видимость понимания; когда же все необходимые вычисления посредством соответствующих манипуляций воспроизводит *человек*, какого-либо понимания в действительности не возникает. На этом основании Серл утверждает, что понимание как свойство мышления не может сводиться исключительно к вычислениям — хотя человек (не знающий китайского) и воспроизводит каждую вычислительную операцию, выполняемую компьютером, он все же совершенно не понимает смысла рассказанной истории. Серл допускает, что возможно осуществить *моделирование* получаемых на выходе результатов понимания (в полном соответствии с точкой зрения $\mathscr{B}$), поскольку он полагает, что это вполне достижимо посредством компьютерного моделирования всей физической активности мозга (чем бы мозг при этом ни занимался) в тот момент, когда его владелец вдруг что-либо

понимает. Однако главный вывод из «китайской комнаты» Джона Серла заключается в том, что сама по себе *модель* в принципе не способна действительно «ощутить» понимание. То есть для любой компьютерной модели *подлинное* понимание остается, в сущности, недостижимым.

Доказательство Серла направлено против точки зрения $\mathscr{A}$ (согласно которой любая «модель» понимания эквивалентна «подлинному» пониманию) и, по замыслу автора, в поддержку точки зрения $\mathscr{B}$ (хотя в той же мере оно поддерживает и $\mathscr{C}$ или $\mathscr{D}$). Оно имеет дело с *пассивным*, *обращенным внутрь*, или *субъективным* аспектами понимания, однако при этом не отрицает возможности моделирования понимания в его *активном*, *обращенном наружу*, или *объективном* аспектах. Сам Серл однажды заявил: «Несомненно, мозг — это цифровой компьютер. Раз кругом одни цифровые компьютеры, значит, и мозг должен быть одним из них»[18]. Отсюда можно заключить, что Серл готов принять возможность полного моделирования работы обладающего сознанием мозга в процессе «понимания», результатом которого оказалась бы полная тождественность внешних проявлений модели и внешних проявлений действительно мыслящего человеческого существа, что соответствует точке зрения $\mathscr{B}$. Мое же исследование призвано показать, что одними лишь внешними проявлениями «понимание» отнюдь не ограничивается, в связи с чем я утверждаю, что невозможно построить достоверную компьютерную модель даже внешних проявлений понимания. Я не привожу здесь аргументацию Серла в подробностях, поскольку точку зрения $\mathscr{C}$ она напрямую не поддерживает (а целью всех наших дискуссий здесь является как раз поддержка $\mathscr{C}$ и ничто иное). Тем не менее, следует отметить, что концепция «китайской комнаты» предоставляет, на мой взгляд, достаточно убедительный аргумент против $\mathscr{A}$, хоть я и не считаю этот аргумент решающим. Более подробное изложение и различные контраргументы представлены в [340], обсуждение — там же и в [203]; см. также [80] и [341]. Мою оценку можно найти в НРК, с. 17–23.

## 1.14. Некоторые проблемы вычислительной модели

Прежде чем перейти к вопросам, отражающим специфические отличия точки зрения $\mathscr{C}$ от $\mathscr{A}$ и $\mathscr{B}$, рассмотрим некоторые другие трудности, с которыми непременно сталкивается любая



попытка объяснить феномен сознания в соответствии с точкой зрения $\mathscr{A}$. Согласно $\mathscr{A}$, для возникновения осознания необходимо лишь простое «выполнение» или *воспроизведение* надлежащих алгоритмов. Что же это означает в действительности? Следует ли под «воспроизведением» понимать, что в соответствии с последовательными шагами алгоритма должны перемещаться с места на место некие физические материальные объекты? Предположим, что эти последовательные шаги записываются строка за строкой в огромную книгу[19]. Являются ли «воспроизведением» действия, посредством которых осуществляется запись или печать этих строк? Достаточно ли для осознания одного лишь статического существования такой книги? А если просто водить пальцем от строчки к строчке — можно ли это считать «воспроизведением»? Или если водить пальцем по символам, набранным шрифтом Брайля? А если проецировать страницы книги одну за другой на экран? Является ли воспроизведением простое *представление* последовательных шагов алгоритма? С другой стороны, необходимо ли, чтобы кто-нибудь проверял, на самом ли деле каждая последующая линия надлежащим образом следует из предыдущей (в соответствии с правилами рассматриваемого алгоритма)? Последнее предположение способно, по крайней мере, разрешить все наши сомнения, поскольку данный процесс должен, по всей видимости, обходиться без участия (сознательного) каких бы то ни было ассистентов. И все же нет совершенно никакой ясности относительно того, какие именно физические действия следует считать действительными исполнителями алгоритма осознания. Быть может, подобные действия не требуются вовсе, и можно, не противореча точке зрения $\mathscr{A}$, утверждать, что для возникновения «осознания» вполне достаточно одного лишь теоретического математического существования соответствующего алгоритма (см. § 1.17).

Как бы то ни было, можно предположить, что, даже согласно $\mathscr{A}$, далеко не *всякий* сложный алгоритм может обусловить возникновение осознания (ощущения осознания). Наверное, для того, чтобы можно было считать состоявшимся сколько-нибудь заметное осознание, алгоритм, судя по всему, должен обладать некоторыми особенными свойствами — такими, например, как «высокоуровневая организация», «универсальность», «самоотносимость», «алгоритмическая простота/сложность»[20] и тому подобными. Кроме того, донельзя скользким представляется во-

прос о том, какие именно свойства алгоритма отвечают в этом случае за различные *qualia* (ощущения), формирующие осознание. Например, какое конкретно вычисление вызывает ощущение «красного»? Какие вычисления дают ощущения «боли», «сладости», «гармоничности», «едкости» и т. д.? Сторонники $\mathcal{A}$ время от времени предпринимают попытки разобраться в подобного рода проблемах (см., например, [81]), однако пока что эти попытки выглядят весьма и весьма неубедительными.

Более того, любое четко определенное и достаточно простое алгоритмическое предположение (подобное всем тем, что до сих пор выдвигались в соответствующих исследованиях) обладает одним существенным недостатком: этот алгоритм можно без особых усилий реализовать на современном электронном компьютере. А между тем, согласно утверждению автора такого предположения, реализация его алгоритма неизбежно вызывает *реальное* ощущение того или иного *qualium*. Мне думается, что даже самому стойкому приверженцу точки зрения $\mathcal{A}$ будет сложно всерьез поверить, что такое вычисление — да и вообще любое вычисление, которое можно запустить на современном компьютере, работа которого основывается на современных представлениях об ИИ, — может *действительно* обусловить мышление хотя бы даже и в самой зачаточной степени. Так что сторонникам подобных предположений остается, по всей видимости, уповать лишь на то, что всеми мыслительными ощущениями мы обязаны не чему иному, как банальной *сложности* сопровождающих деятельность мозга вычислений (выполняющихся в соответствии с упомянутыми предположениями).

В связи с этим возникает еще несколько проблем, которых, насколько мне известно, всерьез пока не касался никто. Если предположить, что необходимым условием сознательной мыслительной деятельности является, главным образом, огромная сложность «соединений», формирующих в мозге сеть из взаимосвязанных нейронов и синапсов, то придется каким-то образом примириться и с тем, что сознание свойственно не всем отделам головного мозга человека в равной степени. Когда термин «мозг» употребляют без каких-либо уточнений, вполне естественно (по крайней мере, для неспециалиста) представлять себе обширные, покрытые извилинами внешние области, образующие так называемую *кору головного мозга*, — состоящий из серого вещества наружный слой *головного мозга*. В коре головного мозга содержится приблизительно сто тысяч миллионов ($10^{11}$) нейронов, что и в самом деле дает ощутимый простор для формирования структур огромной сложности, однако кора — это еще далеко не весь мозг. В задней нижней части мозга находится еще один весьма важный сгусток спутанных нейронов, известный как *мозжечок* (см. рис. 1.6). Мозжечок, судя по всему, неким критическим образом связан с процессом выработки двигательных навыков; его действие можно наблюдать, когда человек овладевает тем или иным движением в совершенстве, т. е. когда движение перестает требовать сознательного обдумывания, как не требует обдумывания, скажем, ходьба. Сначала, когда мы еще только учимся какому-то новому навыку, нам необходимо контролировать свои действия сознательно, и этот контроль, по-видимому, требует существенного участия коры головного мозга. Однако впоследствии, по мере того, как необходимые движения становятся «автоматическими», управление ими постепенно переходит к мозжечку и осуществляется, по большей части, бессознательно. Учитывая, что деятельность мозжечка является, по всей видимости, абсолютно бессознательной, весьма примечателен тот факт,



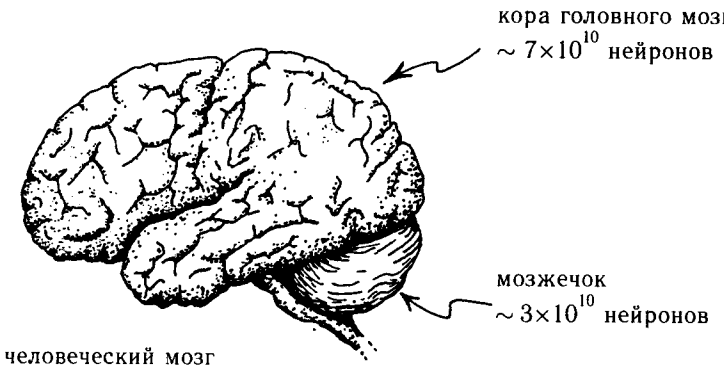



Рис. 1.6. Количество нейронов и нейронных связей в мозжечке совпадает по порядку величины с количеством нейронов и нейронных связей головного мозга. Если основываться лишь на подсчете нейронов и взаимосвязей между ними, то не совсем ясно, почему же деятельность мозжечка абсолютно бессознательна?

что количество нейронов в мозжечке может достигать половины того их количества, что содержится в коре головного мозга. Более того, именно в мозжечке располагаются такие нейроны, как клетки Пуркинье (те самые, что имеют до 80 000 синаптических связей, о чем я уже упоминал в § 1.2), так что общее число связей между нейронами в мозжечке может оказаться ничуть не меньше аналогичного числа в головном мозге. Если необходимым условием возникновения сознания считать одну лишь сложность нейронной сети, то неплохо было бы выяснить, почему же сознание никак, на первый взгляд, не проявляется в деятельности мозжечка. (Несколько дополнительных замечаний на эту тему приведены в § 8.6.)

Разумеется, затронутые в этом разделе проблемы, с которыми приходится иметь дело сторонникам точки зрения $\mathscr{A}$, имеют свои аналоги и применительно к точкам зрения $\mathscr{B}$ и $\mathscr{C}$. Какой бы научной позиции вы ни придерживались, вам в конечном итоге все равно придется как-то решать вопрос о том, что же лежит в основе феномена сознания и как возникают *qualia*. В последних параграфах второй части книги я попытаюсь наметить некоторые пути к пониманию сознания с точки зрения $\mathscr{C}$.

## 1.15. Свидетельствуют ли ограниченные возможности сегодняшнего ИИ в пользу $\mathscr{C}$?

Но почему вдруг $\mathscr{C}$? Чем мы *реально* располагаем, что можно было бы интерпретировать как прямое свидетельство в пользу точки зрения $\mathscr{C}$? Представляет ли $\mathscr{C}$ действительно сколько-нибудь серьезную альтернативу точкам зрения $\mathscr{A}$, $\mathscr{B}$ или даже $\mathscr{D}$? Нам необходимо постараться понять, что именно мы делаем нашим мозгом (или разумом), когда дело доходит до сознательных размышлений; я же попытаюсь убедить читателя в том, что его связанная с сознательным мышлением деятельность весьма отличается (по крайней мере, иногда) от того, что можно реализовать посредством вычислений. Приверженцы точки зрения $\mathscr{A}$, скорее всего, будут утверждать, что мышление осуществляется исключительно посредством «вычислений» в той или иной форме, и никак иначе, — а до тех пор, пока речь идет лишь о внешних проявлениях процесса мышления, с ними согласятся и сторонники $\mathscr{B}$. Что же касается поборников $\mathscr{D}$, то они вполне могли бы



согласиться с $\mathscr{C}$ в том, что деятельность сознания должна быть феноменом невычислимым, однако при этом они будут напрочь отрицать любую возможность объяснения сознания в научных терминах. Таким образом, для поддержания точки зрения $\mathscr{C}$ необходимо найти примеры мыслительной деятельности, не поддающиеся никакому вычислению, и, кроме того, попытаться сообразить, как подобная деятельность может оказаться результатом тех или иных физических процессов. Остаток первой части моей книги будет направлен на достижение первой цели, во второй же части я представлю свои попытки продвинуться по направлению к цели номер два.

Какой же должна быть мыслительная деятельность, чтобы ее невычислимость можно было явственно продемонстрировать? В качестве возможного пути к ответу на этот вопрос можно попытаться рассмотреть современное состояние искусственного интеллекта и постараться понять сильные и слабые стороны систем, управляемых посредством вычислений. Безусловно, сегодняшнее положение дел в области исследований ИИ может и не дать сколько-нибудь четких указаний относительно принципиально возможных достижений будущего. Даже, скажем, через пятьдесят лет ситуация вполне может оказаться совершенно отличной от той, что мы имеем сегодня. Быстрое развитие компьютерных технологий и областей их применения только за *последние* пятьдесят лет привело к чрезвычайно серьезным переменам. Нам, несомненно, следует быть готовыми к значительным переменам и в дальнейшем — переменам, которые, возможно, произойдут с нами очень и очень скоро. И все же в данной книге меня прежде всего будут интересовать не темпы технического развития, а некоторые фундаментальные и *принципиальные* ограничения, которым его достижения неминуемо оказываются подвержены. Эти ограничения останутся в силе независимо от того, на сколько веков вперед мы устремим свой взгляд. Таким образом, свою аргументацию нам следует строить исходя из общих принципов, не предаваясь чрезмерным восторгам по поводу тех или иных сегодняшних достижений. Тем не менее, успехи и неудачи современных исследований искусственного интеллекта вполне могут содержать некоторые полезные для нас ключи, несмотря даже на тот факт, что результаты этих исследований демонстрируют на данный момент лишь очень слабое подобие того, что можно было бы назвать действительно убедительным искусственным интел-

лектом, и это, безусловно, подтвердят даже самые ярые поборники идеи ИИ.

Как ни удивительно, главную неудачу современный искусственный интеллект терпит вовсе не в тех областях, где человеческий разум может вполне самостоятельно продемонстрировать поистине впечатляющую мощь — там, например, где отдельные люди-эксперты способны буквально потрясти всех окружающих какими-то своими специальными познаниями или способностью мгновенно выносить суждения, требующие крайне сложных вычислительных процедур, — а в вещах вполне «обыденных», какие на протяжении большей части своей сознательной жизни проделывают самые заурядные из представителей рода человеческого. Пока что ни один управляемый компьютером робот не может соперничать даже с малым ребенком в таком, например, простейшем деле, как сообразить, что для завершения рисунка необходим цветной карандаш, который валяется на полу в противоположном конце комнаты, после чего подойти к нему, взять и использовать по назначению. Коли уж на то пошло́, даже способности муравья, проявляющиеся в выполнении повседневной муравьиной работы, намного превосходят все то, что можно реализовать с помощью самых сложных современных систем компьютерного управления. А с другой стороны, перед нами имеется поразительный пример *способности* компьютеров к чрезвычайно эффективным действиям — я имею в виду последние работы по созданию шахматных компьютеров. Шахматы, несомненно, представляют собой такой вид деятельности, в котором мощь человеческого интеллекта проявляется особенно ярко, хотя в полной мере эту мощь используют, к сожалению, лишь немногие. И все же современные компьютерные системы играют в шахматы необычайно хорошо и способны выиграть у большинства шахматистов-людей. Даже лучшим из шахматистов приходится сейчас нелегко, и вряд ли им удастся надолго сохранить свое теперешнее превосходство над наиболее продвинутыми компьютерами[21]. Существует еще несколько узких областей, в которых компьютеры могут с успехом (постоянным или переменным) соперничать со специалистами-людьми. Кроме того, необходимо упомянуть и о таких видах интеллектуальной деятельности (например, о прямых численных расчетах), где способности компьютеров значительно превосходят способности людей.



Как бы то ни было, вряд ли можно утверждать, что во всех вышеперечисленных ситуациях компьютер и впрямь *понимает*, что именно он делает. В случае нисходящей организации причина успешной работы системы состоит не в том, что что-то такое понимает сама *система*, а в том, что в управляющую действиями системы программу было изначально заложено понимание, присущее программистам (или экспертам, которые наняли программистов). Что же касается восходящей организации, то не совсем ясно, есть ли здесь вообще необходимость в каком бы то ни было специфическом понимании на системном уровне либо со стороны самого устройства, либо со стороны программистов, за исключением того понимания, которое потребовалось при разработке конкретных алгоритмов, используемых устройством для улучшения качества своей работы, и того понимания, что изначально позволило создать саму концепцию возможности улучшения качества работы системы на основе накапливаемого ею опыта посредством внедрения в нее соответствующей системы обратной связи. Разумеется, не всегда возможно однозначно определить, что же на самом деле означает термин «понимание», вследствие чего кто-то может утверждать, что в *его* (или ее) системе обозначений такие компьютерные системы и в самом деле демонстрируют своего рода «понимание».

Однако разумно ли это? Для иллюстрации отсутствия какого бы то ни было реального понимания у современных компьютеров рассмотрим один занятный пример — шахматную позицию, приведенную на рис. 1.7 (автор: Уильям Хартстон; цитируется по статье Джейн Сеймур и Дэвида Норвуда [342]). В этой позиции черные имеют огромное преимущество по фигурам в виде двух ладьей и слона. И все же белые очень легко избегают поражения, просто делая ходы королем на своей стороне доски. Стена из пешек для черных фигур непреодолима, и черные ладьи или слон не представляют для белых никакой опасности. Это вполне очевидно для любого человека, который в достаточной степени знаком с правилами игры в шахматы. Но когда эту позицию (белые начинают) предложили компьютеру «Deep Thought» — самому мощному на то время шахматному компьютеру, имеющему в своем активе несколько побед над гроссмейстерами-людьми, — он тут же совершил грубейшую ошибку, взяв пешкой черную ладью, что разрушило заслон из пешек и поставило белых в безнадежно проигрышное положение!

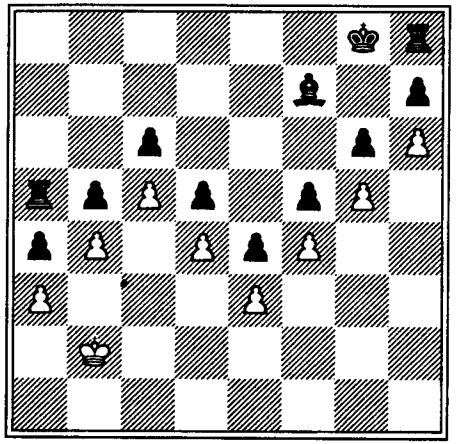


Рис. 1.7. Белые начинают и заканчивают игру вничью — очевидно для человека, а вот «Deep Thought» взял ладью!

Как мог столь искусный шахматист сделать такой очевидно глупый ход? Ответ заключается в следующем: помимо большого количества «позиций из учебника» программа «Deep Thought» содержала лишь инструкции, которые сводились исключительно к вычислению последовательности будущих ходов (на некоторую значительную глубину), позволяющей достичь максимального преимущества по фигурам. Ни на одном из этапов вычислений компьютер не обладал подлинным пониманием не только того, что может ему дать заслон из пешек, но и вообще любого из своих действий.

Любой, кто в достаточной степени представляет себе общий принцип работы компьютера «Deep Thought» или других компьютерных систем для игры в шахматы, не станет удивляться тому, что эта система терпит крах в позициях вроде той, что показана на рис. 1.7. *Мы* не только способны понять в шахматах что-то такое, чего не понимает «Deep Thought»; мы, кроме того, кое-что понимаем и в процедурах (нисходящих), на которых построена вся работа «Deep Thought», то есть мы способны как реально оценить, почему он сделал столь грубую ошибку, так и понять, почему в большинстве других случаев он может играть в шахматы настолько эффективно. Напрашивается, однако, вопрос: сможет



ли «Deep Thought» или иная ИИ-система достичь *когда-нибудь* хоть какого-то подлинного понимания — подобного тому, каким обладаем мы сами — в шахматах или в чем-то еще? Некоторые сторонники ИИ скажут, что для обретения ИИ-системой «подлинного» понимания (что бы это ни значило) ее программа должна задействовать *восходящие* процедуры на гораздо более фундаментальном уровне, нежели это принято в программах теперешних шахматных компьютеров. Соответственно, в такой системе «понимание» развивалось бы постепенно по мере накопления «опыта», а не возникало бы в результате введения каких-то конкретных нисходящих алгоритмических правил. Нисходящие правила, достаточно простые и прозрачные, не способны сами по себе обеспечить вычислительную основу для подлинного понимания, поскольку само понимание этих правил позволяет нам осознать их фундаментальные ограничения.

Этот момент мы более подробно рассмотрим в главах 2 и 3. А что же в самом деле восходящие вычислительные процедуры? Могут ли *они* составить основу для понимания? В главе 3 я приведу рассуждения, доказывающие обратное. Пока же мы можем просто взять на заметку тот факт, что современные компьютерные системы восходящего типа никоим образом не обеспечивают замены подлинному человеческому пониманию ни в одной из важных областей интеллектуальной компетенции, требующих настоящего живого человеческого понимания и интуиции. Такую позицию, я уверен, сегодня разделяют многие. Весьма оптимистичные перспективы[22], время от времени выдвигаемые сторонниками идеи искусственного интеллекта и производителями экспертных систем, пока что в большинстве своем реализованы не были.

Однако в том, что касается возможных результатов развития искусственного интеллекта, мы все еще находимся в самом начале пути. Сторонники ИИ (в форме 𝒜 или 𝒝) уверяют нас, что проявление существенных элементов понимания в поведении их систем с компьютерным управлением — всего лишь вопрос времени и, быть может, некоторых, пусть и значительных, технических усовершенствований. Несколько позднее я попробую поспорить с этим заявлением в более точных терминах, опираясь на то, что некие фундаментальные ограничения присущи любой чисто вычислительной системе, будь она нисходящей или восходящей. Не исключая возможности того, что, будучи достаточ-

но грамотно сконструированной, такая система сможет в течение некоторого продолжительного периода времени поддерживать иллюзию обладания чем-то, подобным пониманию (как это произошло с компьютером «Deep Thought»), я все же утверждаю, что на деле полная ее неспособность к пониманию в общем смысле этого слова непременно в конце концов обнаружится — по крайней мере, в принципе.

Для приведения точных аргументов мне придется обратиться к математике, причем я намерен показать, что к одним лишь вычислениям невозможно свести даже *математическое* понимание. Некоторые защитники ИИ могут счесть это весьма удивительным, ибо они утверждают[23], что те способности, которые сформировались в процессе эволюционного развития человека сравнительно недавно (например, способность выполнять арифметические или алгебраические вычисления), «осваиваются» компьютерами легче всего, и именно в этих областях компьютеры на настоящий момент значительно опережают «человека вычисляющего»; овладение же теми способностями, что развились в начале эволюционного пути — такими, например, как умение ходить или интерпретировать сложные визуальные сцены, — не требует практически никакого труда от человека, тогда как сегодняшние компьютеры даже при всем старании демонстрируют в этом «виде спорта» весьма посредственные результаты. Я рассуждаю несколько иначе. Современный компьютер легко справится с любой сложной деятельностью — будь то математические вычисления, игра в шахматы или выполнение какой-либо работы по дому, — но *лишь при условии*, что эту деятельность можно описать в виде набора четких вычислительных правил; а вот собственно понимание, лежащее в основе этих самых вычислительных правил, оказывается феноменом, для вычисления недоступным.

## 1.16. Доказательство на основании теоремы Гёделя

Как можем мы быть уверены в том, что вышеописанное понимание не может, в сущности, быть сведено к набору вычислительных правил? Несколько позже (в главах 2 и 3) я приведу некоторые очень серьезные доводы в пользу того, что проявления



понимания (по крайней мере, определенных его видов) невозможно достоверно моделировать посредством каких угодно вычислений — ни нисходящего, ни восходящего типа, ни любой из их комбинаций. Таким образом, за реализацию присущей человеку способности к «пониманию» должна отвечать какая-то невычислительная деятельность мозга или разума. Напомним, что термином «невычислительный» в данном контексте (см. § 1.5, § 1.9) мы характеризуем феномен, который невозможно эффективно моделировать с помощью какого угодно компьютера, основанного на логических принципах, общих для всех современных электронных или механических вычислительных устройств. При этом термин «невычислительная активность» вовсе *не предполагает* невозможности описать такую активность научными и, в частности, математическими методами. Он *предполагает* лишь то, что точки зрения $\mathscr{A}$ и $\mathscr{B}$ оказываются не в состоянии объяснить, каким именно образом мы выполняем все те действия, которые представляют собой результат сознательной мыслительной деятельности.

Существует, по меньшей мере, *логическая* возможность того, что обладающий сознанием мозг (или сознательный разум) может функционировать в соответствии с такими невычислительными законами (см. § 1.9). Однако *так* ли это? Представленные в следующей главе (§ 2.5) рассуждения содержат, как мне кажется, весьма четкое доказательство наличия в нашем сознательном мышлении невычислительной составляющей. Основаны эти рассуждения на знаменитой и мощной теореме математической логики, сформулированной великим логиком, чехом по происхождению, Куртом Гёделем. Для моих целей будет вполне достаточно существенно упрощенного варианта этой теоремы, который не потребует от читателя слишком обширных познаний в математике (что касается математики, то я также позаимствую кое-что из одной важной идеи, высказанной несколько позднее Аланом Тьюрингом). Любой достаточно серьезно настроенный читатель без труда разберется в моих рассуждениях. Доказательства гёделевского типа, да еще и примененные в подобном контексте, подвергаются время от времени решительным нападкам[24]. Вследствие этого у некоторых читателей может сложиться впечатление, что мое основанное на теореме Гёделя доказательство было полностью опровергнуто. Должен заметить, что это *далеко* не так. За прошедшие годы действительно выдвигалось мно-

жество контраргументов. Мишенью для многих из них послужило одно из самых первых таких доказательств (направленное в поддержку ментализма и против физикализма), предложенное оксфордским философом Джоном Лукасом [246]. Опираясь на результаты теоремы Гёделя, Лукас доказывал, что мыслительные процессы невозможно воспроизвести вычислительными методами. (Подобные соображения выдвигались и ранее; см., например, [271].) Мое доказательство, пусть и построенное на том же фундаменте, выдержано все же в несколько ином духе, нежели доказательство Лукаса; кроме того, в число моих задач не входила непременная поддержка ментализма. Я думаю, что моя формулировка способна лучше противостоять различным критическим замечаниям, выдвинутым в свое время против доказательства Лукаса, и во многих отношениях выявить их несостоятельность.

Ниже (в главах 2 и 3) мы подробно рассмотрим *все* контраргументы, которые когда-либо попадались мне на глаза. Надеюсь, что мои сопутствующие комментарии не только помогут прояснить некоторые, похоже, широко распространившиеся заблуждения относительно смысла доказательства Гёделя, но и дополнят, по-видимому, неудовлетворительно краткое рассмотрение этого вопроса, предпринятое в НРК. Я намерен показать, что большая часть этих контраргументов произрастает, в сущности, из банальных недоразумений, тогда как остальные, основанные на более или менее осмысленных и требующих детального рассмотрения возражениях, представляют собой, в лучшем случае, не более чем возможные «лазейки» в духе взглядов $\mathscr{A}$ или $\mathscr{B}$; при этом они *не дают* — в чем у нас еще будет возможность убедиться — сколько-нибудь *правдоподобного* объяснения действительным последствиям наличия у нас способности «понимать», да и в любом случае эти лазейки не представляют особой ценности для развития идеи ИИ. Так что тем, кто по-прежнему полагает, что все внешние проявления процессов сознательного мышления *можно* адекватно воспроизвести вычислительными методами, в рамках положений $\mathscr{A}$ или $\mathscr{B}$, я могу лишь порекомендовать повнимательнее следить за предлагаемой ниже аргументацией.

## 1.17. Платонизм или мистицизм?

Критики, впрочем, могут возразить, что отдельные выводы в рамках этого доказательства Гёделя следует рассматривать не



иначе как «мистические», поскольку упомянутое доказательство, судя по всему, вынуждает нас принять либо точку зрения $\mathscr{C}$, либо точку зрения $\mathscr{D}$; подобный взгляд, разумеется, не более приемлем, нежели любая из вышеупомянутых лазеек, полученных из теоремы Гёделя. Что касается $\mathscr{D}$, то здесь я, вообще говоря, полностью с критиками согласен. Мои собственные причины неприятия $\mathscr{D}$ — точки зрения, настаивающей на полном бессилии науки перед тайною разума, — проистекают из осознания того факта, что только благодаря применению научных и, в частности, математических методов был достигнут хоть какой-то реальный прогресс в понимании происходящих в окружающем нас мире процессов. Более того, если мы и располагаем какими-то достоверными сведениями о разуме, то только о том разуме, который тесно связан с конкретным физическим объектом — *мозгом*, — причем различным состояниям разума четко соответствуют различные физические состояния мозга. По всей видимости, с теми или иными специфическими типами физической активности мозга можно ассоциировать и *психические* состояния *сознания*. Если бы не таинственные аспекты сознания, связанные с формированием «осознания» и, быть может, с проявлениями «свободы воли», которые пока что не поддаются физическому описанию, нам бы и в голову не пришло, что для объяснения разума, являющегося по всем признакам продуктом протекающих внутри мозга физических процессов, стандартных научных методов может и не хватить.

С другой стороны, следует понимать, что наука (и, в частности, математика) и сама по себе являет нам мир, исполненный тайн. Чем глубже мы проникаем в процессе научного познания в суть вещей, тем более фундаментальные тайны открываются нашему взору. Быть может, стóит в этой связи упомянуть и о том, что физики, более непосредственно знакомые с головоломной и непостижимой манерой, в какой *реально* проявляет себя материя, склонны видеть мир в менее классически механистическом свете, нежели биологи. В главе 5 мы поговорим о некоторых наиболее таинственных аспектах квантового поведения, обнаруженных относительно недавно. Возможно, для полного «охвата» тайны разума нам придется несколько расширить границы того, что мы в настоящее время называем *наукой*, однако я не вижу причин напрочь отказываться от тех методов, которые так замечательно служили нам до сих пор. Таким образом, если

гёделевские соображения подталкивают нас к принятию точки зрения $\mathscr{C}$ в том или ином ее виде (а я полагаю, что так оно и есть), то нам поневоле придется принять и некоторые другие ее следствия. Иными словами, следуя этим путем, мы приходим, ни много ни мало, к объективному идеализму по *Платону*. Согласно учению Платона, математические концепции и математические истины существуют в их собственном, вполне реальном мире, в котором отсутствует течение времени и который не имеет физического местонахождения. Мир Платона — это идеальный мир совершенных форм, отличный от физического мира, но являющийся основой для его понимания. Он, кроме того, никак не связан с нашими несовершенными мысленными построениями, однако человеческий разум способен получить в некотором смысле непосредственный доступ в это платоново царство благодаря способности «осознавать» математические формы и рассуждать о них. Нашему «платоническому» восприятию, как вскоре выяснится, может иногда поспособствовать вычисление, однако в общем это восприятие вычислением не ограничено. Согласно такому платоническому подходу, именно способность «осознавать» математические концепции дает разуму мощь, далеко превосходящую все, чего можно добиться от устройства, работа которого основывается исключительно на вычислении.

## 1.18. Почему именно математическое понимание?

Все эти благоглупости, конечно, очень (или не очень) замечательны — так, несомненно, уже ворчат иные читатели. Однако какое отношение имеют все эти замысловатые проблемы математики и философии математики к большинству вопросов, непосредственно касающихся, например, искусственного интеллекта? В самом деле, многие философы и поборники ИИ придерживаются достаточно разумного мнения, суть которого сводится к тому, что теорема Гёделя, безусловно, имеет огромное значение в своем исходном контексте, т. е. в области математической логики, однако в отношении ИИ или философии разума актуальность ее, в лучшем случае, весьма и весьма ограничена. В конце концов, не так уж и часто мыслительная деятельность человека оказывается направлена на решение вопросов, относящихся к первоначальной области применимости рассуждений Гёделя — аксиоматическим основам математики. На это возражение я бы ответил так: но ведь практически всегда мыслительная деятельность человека требует участия сознания и понимания. Рассуждение же Гёделя я использую для того, чтобы показать, что человеческое понимание нельзя свести к алгоритмическим процессам. Если мне удастся показать справедливость этого утверждения в *каком-либо* конкретном контексте, то этого будет вполне достаточно. Продемонстрировав, что понимание каких-то математических процедур не поддается описанию с помощью вычислительных методов, мы тем самым докажем, что в нашем разуме происходит-таки *что-то* такое, что невозможно вычислить. А если так, то напрашивается вполне естественный вывод: невычислительная активность должна быть присуща и многим другим аспектам мыслительной деятельности. Вот и все, путь свободен!



Может показаться, что представленное в главе 2 математическое доказательство, устанавливающее необходимую нам форму теоремы Гёделя, не имеет прямого отношения к большинству аспектов сознания. В самом деле: что общего может быть у демонстрации невычислимости феномена понимания на примере определенных типов математических суждений с восприятием, например, красного цвета? Да и в большинстве других аспектов сознания математические соображения, похоже, не играют явно выраженной роли. К примеру, даже математики, как правило, не думают о математике, когда спят и видят сны! Судя по всему, сны видят и собаки, причем есть основания полагать, что они, до некоторой степени, осознают, что видят сон; и я склонен думать, что они наверняка осознают и происходящее с ними во время бодрствования. Однако собаки математикой не занимаются. Бесспорно, математические размышления — далеко не *единственная* деятельность живого организма, требующая участия сознания. Скажем больше: эта деятельность в высшей степени специализирована и характерна лишь для человека. (И даже более того, я встречал циников, которые уверяли меня, что упомянутая деятельность характерна лишь для определенной, чрезвычайно редкой разновидности людей.) Феномен же сознания наблюдается повсеместно и присущ мыслительной деятельности как человека, так и большинства нечеловеческих форм жизни; сознанием, безусловно, в равной степени обладают и люди, далекие от математики, и математики-профессионалы, причем даже тогда, когда они математикой не занимаются (т. е. бóльшую часть своей

жизни). Математическое мышление составляет очень и очень малую область сознательной деятельности вообще, практикует его очень и очень незначительное меньшинство обладающих сознанием существ, да и то на протяжении очень и очень ограниченной части их сознательной жизни.

Почему же в таком случае я решил рассмотреть вопрос сознания прежде всего в математическом контексте? Причина заключается в том, что только в математических рамках мы можем рассчитывать на возможность хоть сколько-нибудь строгой демонстрации *непременной* невычислимости, по крайней мере, *некоторой* части сознательной деятельности. Вопрос вычислимости по самой своей природе является, безусловно, математическим. Нельзя ожидать, что нам удастся дать хоть какое-то «доказательство» невычислимости того или иного процесса, не обратившись при этом к математике. Я хочу убедить читателя в том, что все, что мы делаем нашим мозгом или разумом в процессе понимания *математического* суждения, существенно отличается от того, чего мы можем добиться от какого угодно компьютера; если мне это удастся, то читателю будет намного легче оценить роль невычислительных процессов в сознательном мышлении вообще.

А разве не *очевидно*, возразят мне, что восприятие того же красного цвета никак не может быть вызвано просто выполнением какого бы то ни было вычисления. К чему вообще утруждать себя какими-то ненужными математическими демонстрациями, когда и без того совершенно ясно, что *qualia* — т. е. субъективные ощущения — никак не связаны с вычислениями? Один из ответов заключается в том, что такое доказательство от «очевидного» (как бы благожелательно я ни относился к подобному способу доказательства) применимо только к *пассивным* аспектам сознания. Как и китайскую комнату Серла, его можно представить в качестве аргумента против точки зрения $\mathscr{A}$, а вот между $\mathscr{C}$ и $\mathscr{B}$ разницы для него не существует.

Более того, мне представляется крайне уместным побить функционалистов вместе с их вычислительной моделью (т. е. точкой зрения $\mathscr{A}$), так сказать, на их собственном поле; ведь это именно функционалисты настаивают на том, что все *qualia* на самом деле *должны* быть так или иначе обусловлены банальным выполнением соответствующих вычислений, невзирая на то, сколь невероятной такая картина может показаться на первый взгляд. Ибо, аргументируют они, что же еще можем мы эффективно делать своим мозгом, как не выполнять те или иные вычисления? Для чего вообще нужен мозг, если не в качестве своеобразной системы управления вычислениями — да, чрезвычайно сложными, но все же вычислениями? Какие бы «ощущения осознания» ни пробуждались в нас в результате той или иной функциональной активности мозга, эти ощущения, согласно функционалистской модели, непременно являются результатом некоторой вычислительной процедуры. Функционалисты любят упрекать тех, кто не признает за вычислительной моделью способности объяснить *любые* проявления активности мозга, включая и сознание, в склонности к *мистицизму*. (Надо понимать так, что единственной альтернативой точки зрения $\mathscr{A}$ является $\mathscr{D}$.) Во второй части книги я намерен привести несколько частных предположений относительно того, что еще может вполне эффективно делать мозг, допускающий научное описание. Не стану отрицать, некоторые «конструктивные» моменты моего доказательства являются чисто умозрительными. И все же я полагаю, что мои доводы в пользу невычислимости хотя бы *некоторых* мыслительных процессов весьма убедительны; а для того, чтобы эта убедительность переросла в неотразимость, их следует применить к математическому мышлению.



## 1.19. Какое отношение имеет теорема Гёделя к «бытовым» действиям?

Допустим однако, что мы все уже согласны с тем, что при формировании осознанных математических суждений и получении осознанных же математических решений в нашем мозге действительно происходит что-то невычислимое. Каким образом это поможет нам понять причины ограниченных способностей роботов, которые, как я упоминал ранее, значительно хуже справляются с элементарными, «бытовыми», действиями, нежели со сложными задачами, для выполнения которых требуются высококвалифицированные специалисты-люди? На первый взгляд, создается впечатление, что мои выводы в корне *противоположны* тем, к которым придет всякий здравомыслящий человек, исходя из известных ограничений искусственного интеллекта — по крайней мере, сегодняшних ограничений. Ибо многим почему-то

кажется, что я утверждаю, будто невычислимое поведение должно быть связано скорее с пониманием крайне сложных областей математики, а никак не с обыденным, бытовым поведением. Это не так. Я утверждаю лишь, что *пониманию* сопутствуют невычислимые процессы одинаковой природы, вне зависимости от того, идет ли речь о подлинно математическом восприятии, скажем, бесконечного множества натуральных чисел или всего лишь об осознании того факта, что предметом удлиненной формы можно подпереть открытое окно, о понимании того, какие именно манипуляции следует произвести с куском веревки для того, чтобы привязать или, напротив, отвязать уже привязанное животное, о постижении смысла слов «счастье», «битва» или «завтра» и, наконец, о логическом умозаключении относительно вероятного местонахождения правой ноги Авраама Линкольна, если известно, что левая его нога пребывает в настоящий момент в Вашингтоне, — я привел здесь некоторые из примеров, оказавшихся на удивление мучительными для одной реально существующей ИИ-системы![25] Такого рода невычислимые процессы лежат в основе всякой деятельности, результатом которой является непосредственное осознание чего-либо. Именно это осознание позволяет нам визуализировать геометрию движения деревянного бруска, топологические свойства куска веревки или же «связность» Авраама Линкольна. Оно также позволяет нам получить до некоторой степени прямой доступ к опыту другого человека, с помощью чего мы можем «узнать», что этот другой, скорее всего, подразумевает под такими словами, как «счастье», «битва» и «завтра», несмотря даже на то, что предлагаемые в процессе общения объяснения зачастую оказываются недостаточно адекватными. Передать «смысл» слов от человека к человеку все же возможно, однако не с помощью объяснений различной степени адекватности, а лишь благодаря тому, что собеседник уже, как правило, имеет в сознании некий общий образ возможного смысла этих слов (т. е. «осознает» их), так что даже очень неадекватных объяснений обычно бывает вполне достаточно для того, чтобы человек смог «уловить» верный смысл. Именно наличие такого общего «осознания» делает возможным общение между людьми. И именно этот факт ставит неразумного, управляемого компьютером робота в крайне невыгодное положение. (В самом деле, уже самый *смысл* понятия «смысл слова» изначально воспринимается нами как нечто само собой разумеющееся, и поэтому совершенно непонятно, каким образом *такое* понятие можно сколько-нибудь адекватно описать нашему неразумному роботу.)



Смысл можно передать лишь от человека к человеку, потому что все люди имеют схожий жизненный опыт или аналогичное внутреннее ощущение «природы вещей». Можно представить «жизненный опыт» в виде своеобразного хранилища, в которое складывается память обо всем, что происходит с человеком в течение жизни, и предположить, что нашего робота не так уж и сложно таким хранилищем оснастить. Однако я утверждаю, что это не так; ключевым моментом здесь является то, что рассматриваемый субъект, будь то человек или робот, должен свой жизненный опыт *осознавать*.

Что же заставляет меня утверждать, будто упомянутое осознание, что бы оно из себя ни представляло, должно быть невычислимым — иначе говоря, таким, что его не сможет ни достичь, ни хотя бы *воспроизвести* ни один робот, управляемый компьютером, построенным исключительно на базе стандартных логических концепций машины Тьюринга (или эквивалентной ей) нисходящего либо восходящего типа? Именно здесь и играют решающую роль гёделевские соображения. Вряд ли мы в настоящее время можем многое сказать об «осознании», например, красного цвета; а вот относительно осознания бесконечности множества натуральных чисел кое-что определенное нам таки известно. Это такое «осознание», благодаря которому ребенок «знает», что означают слова «ноль», «один», «два», «три», «четыре» и т. д. и что следует понимать под бесконечностью этой последовательности, хотя объяснения ему были даны до нелепости ограниченные и, на первый взгляд, к делу почти не относящиеся, на примере нескольких бананов и апельсинов. Из таких частных примеров ребенок и в самом деле способен вывести абстрактное понятие числа «три». Более того, он также оказывается в состоянии понять, что это понятие является лишь звеном в бесконечной цепочке похожих понятий («четыре», «пять», «шесть» и т. д.). В некотором платоническом смысле ребенок изначально «знает», что такое натуральные числа.

Возможно, кто-то усмотрит здесь некий налет мистики, однако в действительности мистика здесь не при чем. Для понимания последующих рассуждений крайне важно отличать такое платоническое знание от мистицизма. Понятия, «известные» нам в платоническом смысле, суть вещи для нас «очевидные»: вещи,

которые сводятся к воспринятому когда-то «здравому смыслу», — при этом мы не можем охарактеризовать эти понятия во всей их полноте посредством вычислительных правил. Действительно — и это станет ясно из дальнейших рассуждений, связанных с доказательством Гёделя, — не существует способа целиком и полностью охарактеризовать свойства натуральных чисел на основе лишь таких правил. А как же тогда описания числа через яблоки или бананы дают ребенку понять, что означают слова «три дня», и откуда ему знать, что смысл абстрактного понятия числа «три» здесь совершенно тот же, что и в словах «три апельсина»? Разумеется, такое понимание иногда приходит к ребенку далеко не сразу, и на первых порах он, бывает, ошибается, однако суть не в этом. Суть в том, что подобное осознание вообще возможно. Абстрактное понятие числа «три», равно как и представление о том, что существует бесконечная последовательность аналогичных понятий — собственно последовательность натуральных чисел, — и в самом деле вполне доступно человеческому пониманию, однако, повторяю, лишь через осознание.

Я утверждаю, что точно так же мы не пользуемся вычислительными правилами при визуализации движений деревянного бруска, куска веревки или Авраама Линкольна. Вообще говоря, существуют весьма эффективные компьютерные модели движения твердого тела — например, деревянного бруска. С их помощью можно осуществлять моделирование такого движения с точностью и достоверностью, обычно недостижимыми при непосредственной визуализации. Аналогично, вычислительными методами можно моделировать и движение веревки или струны, хотя такое моделирование почему-то оказывается несколько более сложным по сравнению с моделированием движения твердого тела. (Отчасти это связано с тем, что для описания положения «математической струны» необходимо определить бесконечно много параметров, тогда как положение твердого тела описывается всего шестью.) Существуют компьютерные алгоритмы для определения «заузленности» веревки, однако они в корне отличаются от алгоритмов, описывающих движение твердого тела (и не очень эффективны в вычислительном отношении). Любое воспроизведение с помощью компьютера внешнего облика Авраама Линкольна, безусловно, представляет собой еще более сложную задачу. Во всяком случае, дело не в том, что визуализация чего-либо



человеком «лучше» или «хуже» компьютерного моделирования, просто это вещи совершенно *различные*.

Важный момент, как мне кажется, заключается в том, что визуализация содержит некий элемент оценки того, что человек видит, то есть сопровождается *пониманием*. Чтобы проиллюстрировать, что я имею в виду, давайте рассмотрим одно элементарное арифметическое правило, а именно: для любых двух натуральных чисел (т. е. неотрицательных целых чисел 0, 1, 2, 3, 4, ...) $a$ и $b$ справедливо следующее равенство:

$$a \times b = b \times a.$$

Следует пояснить, что это высказывание не является пустым, хотя части уравнения и имеют различный смысл. Запись $a \times b$ слева означает совокупность $a$ групп по $b$ объектов в каждой; $b \times a$ справа — $b$ групп по $a$ объектов в каждой. В частном случае, например, при $a = 3$ и $b = 5$, запись $a \times b$ можно представить следующим рядом точек:

$$(\bullet\bullet\bullet\bullet\bullet)(\bullet\bullet\bullet\bullet\bullet)(\bullet\bullet\bullet\bullet\bullet),$$

в то время как для $b \times a$ имеем

$$(\bullet\bullet\bullet)(\bullet\bullet\bullet)(\bullet\bullet\bullet)(\bullet\bullet\bullet)(\bullet\bullet\bullet).$$

Общее число точек в каждом случае одинаково, следовательно, справедливо равенство $3 \times 5 = 5 \times 3$.

В истинности этого равенства можно удостовериться, представив зрительно матрицу

$$\begin{array}{ccccc} \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \end{array}$$

Читая матрицу по строкам, можно сказать, что в ней три строки, каждая из которых содержит по пять точек, что соответствует числу $3 \times 5$. Однако если эту же матрицу прочесть по столбцам, то получится пять столбцов по три точки в каждом, что соответствует числу $5 \times 3$. Равенство этих чисел очевидно, поскольку речь в каждом случае идет об одной и той же прямоугольной матрице, просто мы ее по-разному читаем. (Есть и альтернативный вариант: мы можем мысленно повернуть изображение на прямой угол

и убедиться в том, что матрица, соответствующая числу $5 \times 3$, содержит то же количество элементов, что и матрица, соответствующая числу $3 \times 5$.)

Важный момент описанной визуализации заключается в том, что она непосредственно дает нам нечто гораздо более общее, чем просто частное численное равенство $3 \times 5 = 5 \times 3$. Иными словами, в конкретных числовых значениях $a = 3$ и $b = 5$, участвующих в данной процедуре, нет ничего особенного. Полученное правило будет применимо, даже если, скажем, $a = 79\,797\,000\,222$, а $b = 50\,000\,123\,555$, и мы с уверенностью можем утверждать, что

$$79\,797\,000\,222 \times 50\,000\,123\,555 = 50\,000\,123\,555 \times 79\,797\,000\,222,$$

несмотря на то, что у нас нет ни малейшей возможности сколько-нибудь точно представить себе визуально прямоугольную матрицу такого размера (да и ни один современный компьютер не сможет перечислить все ее элементы). Мы вполне можем заключить, что вышеприведенное равенство должно быть истинным — или что истинным должно быть равенство общего вида[6] $a \times b = b \times a$ — на основании, в сущности, той же самой визуализации, которую мы применяли для конкретного случая $3 \times 5 = 5 \times 3$. Нужно просто несколько «размыть» мысленно действительное количество строк и столбцов рассматриваемой матрицы, и равенство становится очевидным.

Я вовсе не хочу сказать, что все математические отношения можно с помощью верной визуализации непосредственно постигать как «очевидные», или же что их просто можно в любом случае постичь каким-то иным способом, основанным непосредственно на интуиции. Это далеко не так. Для уверенного понимания некоторых математических отношений необходимо строить весьма длинные цепочки умозаключений. Цель математического доказательства, по сути дела, в этом и заключается: мы строим цепочки умозаключений таким образом, чтобы на *каждом этапе* получать утверждение, допускающее «очевидное» понимание. Как следствие, конечной точкой умозаключения должно

[6] Необходимо отметить, что это равенство *не является истинным* для различных странных «чисел», встречающихся порой в математике, — например, для трансфинитных чисел, о которых упоминается в пояснении к **Q19**, § 2.10. Однако для *натуральных* чисел, о которых здесь, собственно, и идет речь, оно всегда справедливо.



оказаться суждение, которое необходимо принимать как *истинное*, пусть даже оно само по себе вовсе и не очевидно.

Кое-кто, наверное, уже вообразил, что в таком случае можно раз и навсегда составить список всех «возможных» этапов умозаключений и тогда всякое доказательство можно будет свести к вычислению, т. е. к простым механическим манипуляциям полученными очевидными этапами. Доказательство Гёделя (§ 2.5) как раз и демонстрирует невозможность реализации такой процедуры. Нельзя совершенно избавиться от необходимости в *новых* «очевидно понимаемых» отношениях. Таким образом, математическое понимание никоим образом не сводится к бездумному вычислению.

## 1.20. Мысленная визуализация и виртуальная реальность

Интуитивные математические процедуры, описанные в § 1.19, имеют весьма ярко выраженный специфический геометрический характер. В математических доказательствах применяются и многие другие типы интуитивных процедур, причем некоторые из них весьма далеки от «геометричности». Однако, как показывает практика, геометрические интуитивные представления чаще всего дают более глубокое математическое понимание. Полагаю, было бы весьма полезно выяснить, какие же именно физические процессы происходят в нашем мозге, когда мы визуализируем что-либо геометрически. Начнем хотя бы с того, что никакой логической необходимости в том, чтобы непосредственным результатом этих процессов было «геометрическое отражение» визуализируемого объекта, по сути дела, не существует. Как мы увидим далее, здесь может получиться нечто совсем иное.

Здесь уместно провести аналогию с феноменом, именуемым «виртуальной реальностью». Феномен этот, согласно распространенному мнению, имеет самое прямое отношение к теме «визуализации». Методы виртуальной реальности[(26)] позволяют создать компьютерную модель какой-либо не существующей в природе структуры, — например, здания на стадии архитектурного проекта, — затем модель проецируется в глаз наблюдателя-человека, который, предположительно, воспринимает ее как «реальное» здание. Совершая движения глазами, головой или, мо-

жет быть, ногами, словно прогуливаясь вокруг демонстрируемого ему здания, наблюдатель может разглядывать его с разных сторон — точно так же, как если бы здание действительно было реальным (см. рис. 1.8). Согласно некоторым предположениям[27], выполняемые мозгом в процессе сознательной визуализации операции (какой бы ни была их истинная природа) аналогичны вычислениям, производимым при построении такой виртуальной модели. В самом деле, мысленно осматривая какую-то *реально* существующую неподвижную структуру, человек, по всей видимости, создает в уме некую модель, которая остается неизменной, несмотря на постоянные движения его головы, глаз и тела, приводящие к непрерывной смене образов, возникающих на сетчатке его глаз. Такие поправки на движения тела играют весьма существенную роль при построении виртуальной реальности, и высказывались предположения в том смысле, что нечто подобное должно происходить и при создании «мысленных моделей», представляющих собой результаты актов визуализации. Такие вычисления, разумеется, вовсе не обязаны иметь целью воспроизведение реальной геометрической структуры моделируемой конструкции (или ее «отражение»). Сторонникам точки зрения $\mathscr{A}$ в таком случае пришлось бы рассматривать сознательную визуализацию как результат своего рода численного моделирования окружающего мира в голове человека. Я же полагаю, что всякий раз, когда мы сознательно воспринимаем ту или иную визуальную сцену, сопровождающее этот процесс *понимание* представляет собой нечто, существенно отличное от моделирования мира методами вычислительного характера.

Можно также предположить, что внутри мозга функционирует нечто вроде «аналогового компьютера», в котором моделирование внешнего мира реализуется не с помощью цифровых вычислений, как в современных электронных компьютерах, а с помощью некоторой внутренней структуры, физическое поведение которой каким-то однозначным образом отражает поведение моделируемой внешней системы. Допустим, например, что нам необходимо аналоговое устройство для моделирования движений некоторого внешнего твердого тела. Для создания такого устройства мы, очевидно, воспользуемся весьма простым и естественным способом. Мы отыщем внутри системы реальное физическое тело той же формы (но меньшего размера), что и моделируемый внешний объект; я, разумеется, ни в коем случае не утверждаю,



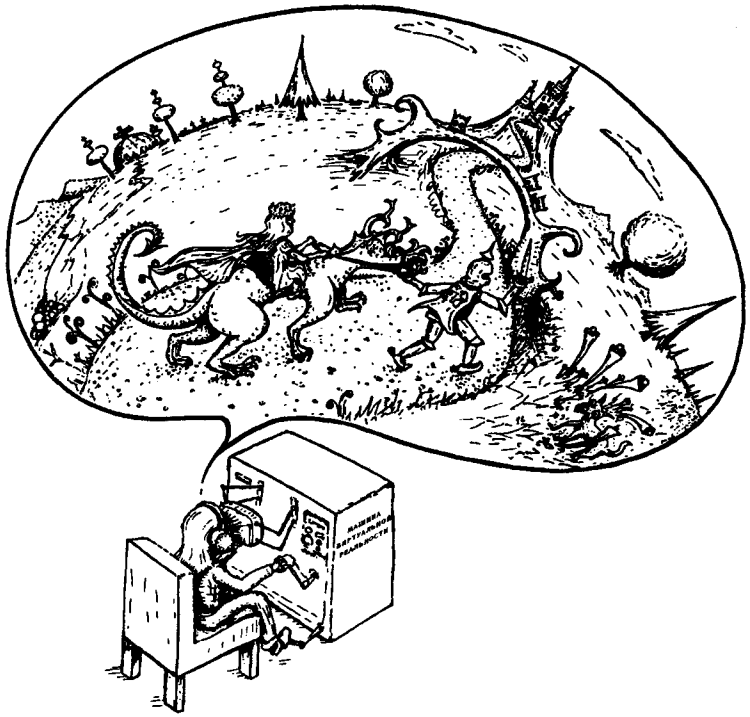


Рис. 1.8. Виртуальная реальность. В результате определенных вычислений в сознании человека возникает трехмерный воображаемый мир, должным образом реагирующий на движения головы и тела наблюдателя.

что данная конкретная модель имеет какое бы то ни было прямое отношение к тому, что происходит внутри мозга. Движения упомянутого «внутреннего» тела можно рассматривать с разных сторон, т. е. в том, что касается внешних проявлений, аналоговая модель оказывается очень похожа на модель, полученную с помощью вычислительных методов. Можно даже создать на основе такой модели систему «виртуальной реальности», в которой вместо целиком вычислительной модели рассматриваемой структуры будет действовать ее реальная физическая модель, отличающаяся от моделируемого «реального» объекта только размерами.

В общем случае аналоговое моделирование вовсе не обязано быть столь прямолинейным и примитивным. Вместо физического расстояния можно использовать в качестве параметра, например, электрический потенциал и т. п. Следует только удостовериться в том, что физические законы, управляющие внутренней структурой, в точности совпадают с физическими законами, которым подчиняется внешняя, моделируемая, структура. При этом нет никакой необходимости в том, чтобы внутренняя структура была *похожа* на внешнюю («отражала» ее) каким-либо очевидным образом.

Способны ли аналоговые устройства достичь результатов, недоступных для чисто вычислительного моделирования? Как уже упоминалось в § 1.8, современная физика не дает никаких оснований полагать, что с помощью аналогового моделирования можно добиться чего-то такого, что принципиально неосуществимо при моделировании цифровом. Иными словами, если мы допускаем, что построение мысленных образов обусловлено какими-то невычислимыми процессами, то это означает, что объяснение данному феномену следует искать за пределами известной нам физики.

## 1.21. Является ли невычислимым математическое воображение?

Говоря о мысленной визуализации, мы ни разу не указали явно на невозможность воспроизведения этого процесса вычислительным путем. Даже если визуализация действительно осуществляется посредством какой-то внутренней аналоговой системы, что мешает нам предположить, что должна существовать, по крайней мере, возможность *смоделировать* поведение такого аналогового устройства?

Дело в том, что «предметом» рассматриваемой выше «визуализации» является «визуальное» в буквальном смысле этого слова, т. е. мысленные образы, соответствующие, как нам представляется, сигналам, поступающим в мозг от глаз. В общем же случае мысленные образы вовсе не обязательно носят такой буквально «визуальный» характер — например, те, что возникают, когда мы понимаем смысл какого-то абстрактного слова или припоминаем музыкальную фразу. Согласитесь, что мысленные об-



разы человека, слепого от рождения, вряд ли могут иметь прямое отношение к сигналам, которые его мозг получает от глаз. Иными словами, под «визуализацией» мы будем в дальнейшем подразумевать скорее процессы, связанные с «осознанием» вообще, нежели те, что имеют непосредственное отношение к системе органов зрения. Честно говоря, мне не известен ни один довод, непосредственно указывающий на вычислительную (или какую-либо иную) природу нашей способности к визуализации именно в буквальном смысле этого слова. Моя же убежденность в том, что процессы «буквальной» визуализации действительно являются невычислимыми, проистекает из явно невычислительного характера *других* видов осознания. Не совсем понятно, каким образом можно произвести прямое доказательство невычислимости исключительно для геометрической визуализации, однако если бы удалось убедительно доказать невычислимость *хотя бы некоторых* форм осмысленного осознания, то такое доказательство дало бы, по меньшей мере, серьезные основания полагать, что вид осознания, ответственный за геометрическую визуализацию, также должен иметь невычислительный характер. По-видимому, нет особой необходимости проводить четкую границу между различными проявлениями феномена сознательного понимания.

Переходя от общего к частному, я утверждаю, что наше понимание, например, свойств натуральных чисел (0, 1, 2, 3, 4, ...) носит *явно* невычислительный характер. (Можно даже сказать, что само понятие натурального числа и есть, в некотором смысле, форма *не*геометрической «визуализации».) В § 2.5, воспользовавшись упрощенным вариантом теоремы Гёделя (см. пояснение к возражению **Q16**), я покажу, что это понимание невозможно описать каким бы то ни было конечным набором правил, а значит, невозможно и воспроизвести с помощью вычислительных методов. Время от времени нас радуют сообщениями о том, что ту или иную компьютерную систему «обучили» «пониманию» концепции натурального числа[28]. Однако, как мы вскоре увидим, этого просто не может быть. Именно *осознание* того, что в действительности может означать слово «число», дает нам возможность верно понять заключенную в нем идею. А располагая верным пониманием, мы — по крайней мере, в принципе — можем давать верные ответы на целый ряд вопросов о числах, буде нам таковые зададут, в то время как ни один конечный набор правил этого обеспечить не в состоянии. Имея в своем распоряжении одни

только правила при полном отсутствии непосредственного осознания, управляемый компьютером робот (такой, например, как «Deep Thought»; см. § 1.15) неизбежно окажется лишен тех способностей, в которых ни один из людей никаких ограничений не испытывает; хотя если снабдить робота достаточно умными правилами поведения, то он, возможно, поразит наше воображение выдающимися интеллектуальными подвигами, многие из которых далеко превзойдут способности обычного человека в каких-то конкретных, достаточно узкоспециальных областях. Возможно даже, что ему удастся на некоторое время одурачить нас, и мы поверим, что и он способен на осознание.

Следует отметить, что всякий раз, как мы получаем *действительно* эффективную цифровую (или аналоговую) компьютерную модель какой-либо внешней системы, это почти всегда происходит благодаря глубокому пониманию человеком тех или иных основополагающих математических идей. Взять хотя бы цифровую модель геометрического движения твердого тела. Выполняемые при таком моделировании вычисления опираются, главным образом, на открытия великих мыслителей семнадцатого века — таких, например, как французские математики Декарт, Ферма и Дезарг, — которым мы обязаны идеями системы координат и проективной геометрии. Существуют и модели, описывающие движение куска веревки или струны. Как выясняется, геометрические идеи, необходимые для понимания особенностей поведения струны — ее так называемой «заузленности», — весьма сложны и относительно молоды. Большинство фундаментальных открытий в этой области были сделаны только в двадцатом веке. Каждый из нас без особого труда способен экспериментальным путем — т. е. посредством несложных манипуляций руками и приложения некоторого здравого смысла — убедиться в наличии либо отсутствии на замкнутой, но спутанной веревочной петле узлов; вычислительные же алгоритмы для достижения того же результата оказываются на удивление сложными и малоэффективными.

Таким образом, эффективное цифровое моделирование таких процессов является в основе своей нисходящим и во многом определяется пониманием и интуитивными прозрениями человека. Вероятность того, что в человеческом мозге при визуализации происходит нечто подобное, очень и очень невелика. Более правдоподобным представляется предположение о том, что существенный вклад в этот процесс вносят те или иные восходящие процедуры, а воспроизводимые в результате «визуальные образы» требуют предварительного накопления немалого «опыта». Я, впрочем, не слышал о сколько-нибудь серьезных исследованиях этого вопроса именно с точки зрения восходящих процедур (например, о разработках искусственных нейронных сетей). По всей видимости, подход, *целиком* основанный на процедурах восходящего типа, даст весьма скудные результаты. Сомневаюсь, что можно построить более или менее удачную модель геометрического движения твердого тела или топологических особенностей движения куска струны при отсутствии подлинного понимания обусловливающих эти движения законов.



Какие же физические процессы следует считать ответственными за осознание — за осознание, которое, судя по всему, необходимо для всякого подлинного понимания? Действительно ли оно не допускает численного моделирования, как того требует точка зрения $\mathscr{C}$? Можно ли, в таком случае, надеяться на какое бы то ни было постижение этого предполагаемого физического процесса — хотя бы в принципе? Думаю, что можно, и более чем уверен, что точка зрения $\mathscr{C}$ представляет собой подлинно научное допущение — просто нужно приготовиться к тому, что наши научные критерии и методы, возможно, претерпят не слишком явные, но весьма существенные изменения. Нужно быть готовым к тому, что объекты наших исследований будут принимать самые неожиданные формы и возникать в таких областях подлинно научного знания, которые, на первый взгляд, никакого отношения к делу не имеют. Читателя, который намерен продолжить чтение этой книги, я прошу сохранять открытость восприятия и вместе с тем внимательно следить за рассуждениями и представляемыми научными свидетельствами, даже если они вдруг покажутся ему несколько сомнительными с точки зрения здравого смысла. Будьте готовы немного поразмыслить над предлагаемыми доводами, а я, в свою очередь, приложу все усилия к изложению их в максимально доступном виде. Уверен, что, настроившись подобным образом, мы с вами преодолеем все преграды.

В оставшихся главах первой части я не буду касаться физики и возможных видов биологической активности, которые способны обусловить невычислимость, требуемую точкой зрения $\mathscr{C}$. Этими предметами мы займемся во второй части книги. Для начала нам предстоит решить вопрос об общей целесообразности по-

исков невычислимых процессов. Пока что вся целесообразность проистекает лишь из моей уверенности в том, что при сознательном понимании мы действительно выполняем какие-то невычислимые операции. Эту уверенность необходимо обосновать, для чего нам придется обратиться к математике.

## Примечания

1. См., в частности, [162], [263], [267].
2. Моравек [267] основывает свои доводы в пользу такого срока на том, какая, по его мнению, часть коры головного мозга успешно реализована в виде модели (речь, в основном, идет о нейронах, расположенных в сетчатке), и на оценке темпов развития компьютерной технологии в ближайшем будущем. Любопытно, что к началу 1994 года он своего мнения не изменил; см. [268].
3. Эти четыре точки зрения были подробно описаны, например, в [215], с. 252 (следует, впрочем, отметить, что условие, называемое автором статьи «тезисом Черча–Тьюринга», является, по своей сути, скорее «тезисом Тьюринга» (в том смысле, в каком я употребляю этот термин в § 1.6), нежели «тезисом Черча»).
4. Например, Д. Деннет, Д. Хофштадтер, М. Мински, Х. Моравек, Г. Саймон; подробнее о терминах можно прочесть в [340], [243].
5. См. [267].
6. [369]; см. также НРК, с. 5–14.
7. См. [340], [341].
8. Вопрос осложняется тем, что современная физика рассматривает, по большей части, *непрерывные*, а не дискретные (цифровые) процессы. Самый *смысл* термина «вычислимость» в данном контексте можно трактовать по-разному. С некоторыми рассуждениями на данную тему можно ознакомиться в [312], [346], [313], [314], [315], [316], [29], [327], [328]. К этому вопросу я еще вернусь в § 1.8.
9. Этой замечательной фразой я обязан диктору BBC Radio 4, ведущему программу «Мысль дня».
10. Исследования в области создания ИИ начались в 1950-е годы с весьма успешного применения сравнительно элементарных нисходящих процедур (например, Грей Уолтер, 1953). Распознающий образы «перцептрон» Фрэнка Розенблатта [323] стал в 1959 году первым удачным «свя́зным» устройством (искусственной нейронной сетью), вызвав тем самым значительный интерес к схемам восходящего типа. В 1969 году Марвин Мински и Сеймур Пейперт указали на некоторые существенные ограничения, присущие данному типу



восходящей организации (см. [264]). Способ обойти эти ограничения предложил некоторое время спустя Хопфилд [207], и в настоящий момент искусственными устройствами, функционирующими по типу нейронной сети, активно занимаются ученые всего мира. (О применении таких устройств, например, в физике высоких энергий см. [19] и [142].) Что касается ИИ нисходящего типа, то здесь важными вехами стали работы Джона Маккарти [248] и Алана Ньюэлла в сотрудничестве с Гербертом Саймоном [272]. Впечатляющее изложение истории исследований проблемы ИИ можно найти в [124]. Из прочей литературы порекомендую [175], [15] (относительно недавние размышления о процедурах и перспективах ИИ); [98] (классическая критика идеи ИИ); [140] (свежий взгляд на проблему от пионера ИИ); также см. статьи в сборниках [40] и [221].
11. Описание $\lambda$-исчисления см. в [52] и [223].
12. Из различных публикаций, посвященных данной проблематике, могу порекомендовать, например, [312], [346], [316], [29]. Вопрос о функционировании мозга в связи с упомянутыми проблемами рассмотрен, в частности, в [326].
13. В действительности Роберт Бергер доказал, что общего алгоритмического решения не имеет лишь задача о замощении плоскости плитками *Вана*. Плитки Вана (названные так в честь математика Хао Вана) представляют собой единичные квадраты с окрашенными краями; при замощении цвета́ соседних плиток должны совпадать, сами же плитки при этом нельзя ни вращать, ни переворачивать. Впрочем, для любого набора плиток Вана несложно составить такой набор полиомино, которым можно будет замостить плоскость тогда и только тогда, когда ее можно замостить соответствующим набором плиток Вана. Таким образом, неразрешимость вычислительными методами задачи о замощении плоскости набором полиомино непосредственно следует из неразрешимости задачи о замощении плоскости набором плиток Вана.

    В связи с задачей о замощении плоскости полиомино следует отметить, что если каким-либо набором полиомино *не удается* замостить плоскость, то этот факт вполне *возможно* установить вычислительным путем (точно так же, как мы можем предсказать остановку машины Тьюринга или убедиться в наличии решения у системы диофантовых уравнений), нужно лишь попытаться замостить плитками данного набора квадратную область размера $n \times n$ (последовательно увеличивая значение $n$); замостить всю плоскость не удастся уже при некотором *конечном* значении $n$. Алгоритмическим путем невозможно установить как раз те случаи, когда данным набором плиток можно-таки *замостить* плоскость.

14. О некоторых чересчур оптимистичных прогнозах относительно ИИ можно прочесть в [124].
15. Своим знакомством с этими вопросами я обязан очень многим людям, среди которых хочу особо поблагодарить Ли Левингера. Замечательное исследование связи современной физики и вычислительных методов с проблемами человеческого поведения можно найти в книге [200].
16. Сломен [344], например, пеняет мне на то, что в НРК я слишком часто прибегаю к такому неопределенному термину, как «сознание», в то время как сам он весьма свободно оперирует еще более неопределенным (на мой взгляд) термином «разум»!
17. См. [340], [341].
18. См. статью Серла [340] (ее также можно найти в сборнике [203], с. 372). Мне, правда, не совсем ясно, к какой точке зрения Серл склонился бы сейчас, к $\mathscr{B}$ или все же к $\mathscr{C}$.
19. Занимательное рассмотрение подобного предположения представлено в [202]; см. также НРК, с. 21–22.
20. Суть понятия «алгоритмической сложности» доступным языком изложена в [45].
21. См. [208].
22. См. [124].
23. См., например, [268].
24. О доказательстве Лукаса см. [320], [345], [24], [163], [164], [236], [237], [202], [37]; см. также [247]. Что касается моей версии, кратко представленной в НРК, с. 416–418, то где только ее не критиковали: см., в особенности, [344] и многочисленные статьи в *Behavioral and Brain Sciences*: [36], [42], [46], [73], [74], [80], [97], [154], [199], [220], [251], [250], [253], [269], [307], [324], [366], [386]; мои ответы на критику см. в [292], [298] и [178]; см. также [95], [294].
25. Примеры взяты из какой-то английской телевизионной программы; возможно, из «Машины мечты» (*The Dream Machine*, декабрь 1991 г.) — четвертой из цикла программ BBC «Мыслящая машина» (*The Thinking Machine*). О последних достижениях в области «искусственного понимания», а в особенности о захватывающем проекте Дугласа Лената «CYC» можно прочесть в [124].
26. Весьма живо и популярно все это описано в [389].
27. Подобное предположение выдвинул, например, Ричард Доукинс в своих «Рождественских лекциях» (BBC, 1992 г.).
28. См., например, рассказ Фридмена [124] о работе Лената и других исследователей в этом направлении.

# 2

# ГЁДЕЛЕВСКОЕ ДОКАЗАТЕЛЬСТВО

## 2.1. Теорема Гёделя и машины Тьюринга

В наиболее чистом виде мыслительные процессы проявляются в сфере математики. Если же мышление сводится к выполнению тех или иных вычислений, то *математическое* мышление, по всей видимости, должно обладать этим свойством в наибольшей степени. Однако, как это ни удивительно, в действительности все происходит с точностью до наоборот. Именно математика дает нам самое явное свидетельство тому, что процессы сознательного мышления включают в себя нечто, не доступное вычислению. Возможно, это покажется парадоксальным, однако для того, чтобы двигаться дальше, нам придется пока с этим парадоксом как-то примириться.

Прежде чем мы начнем, мне бы хотелось хоть как-то успокоить читателя в отношении математических формул, которые встретятся нам в нескольких последующих разделах (§§ 2.2–2.5), хотя надо признать, что страхи его не лишены оснований: ведь нам предстоит в какой-то мере уяснить для себя смысл и следствия ни много ни мало самой важной теоремы математической логики — знаменитой теоремы Курта Гёделя. Я привожу здесь очень и очень упрощенный вариант этой теоремы, опираясь, в частности, на несколько более поздние идеи Алана Тьюринга. Мы не будем пользоваться каким бы то ни было математическим формализмом, за исключением простейшей арифметики. Представленное доказательство, вероятно, будет кое-где несколько путаным, однако *всего лишь* путаным, а ни в коем случае не

«сложным» в смысле необходимости каких-то предварительных познаний в математике. Воспринимайте доказательство в любом удобном для вас темпе и не стесняйтесь перечитывать его столько раз, сколько захочется. В дальнейшем (§§ 2.6–2.10) мы рассмотрим некоторые более специфические соображения, лежащие в основе теоремы Гёделя, однако читатель, не интересующийся подобными вопросами, может эти разделы пропустить без ущерба для понимания.

Так что же такое теорема Гёделя? В 1930 году на конференции в Кёнигсберге блестящий молодой математик Курт Гёдель произвел немалое впечатление на ведущих математиков и логиков со всего мира, представив их вниманию теорему, которая впоследствии получила его имя. Ее довольно быстро признали в качестве фундаментального вклада в основы математики — быть может, наиболее фундаментального из всех возможных, — я же, в свою очередь, утверждаю, что своей теоремой Гёдель также положил начало важнейшему этапу развития философии разума.

Среди положений, которые со всей неоспоримостью доказал Гёдель, имеется следующее: нельзя создать такую *формальную систему* логически обоснованных математических правил доказательства, которой было бы достаточно, хотя бы в принципе, для доказательства всех истинных теорем элементарной арифметики. Уже и это само по себе в высшей степени удивительно, однако это еще не все. Многое говорит за то, что результаты Гёделя демонстрируют нечто большее, — а именно, доказывают, что способность человека к пониманию и постижению сути вещей невозможно свести к какому бы то ни было набору вычислительных правил. Иными словами, нельзя создать такую систему правил, которая оказалась бы достаточной для доказательства даже тех арифметических положений, истинность которых, в принципе, доступна для человека с его интуицией и способностью к пониманию, а это означает, что человеческие интуицию и понимание невозможно свести к какому бы то ни было набору правил. Последующие мои рассуждения отчасти имеют целью убедить читателя в том, что вышеприведенное утверждение действительно следует из теоремы Гёделя; более того, именно на теореме Гёделя основывается мое доказательство неизбежности наличия в человеческом мышлении составляющей, которую никогда не удастся воспроизвести с помощью компьютера (в том смысле, который мы вкладываем в этот термин сегодня).



Думаю, нет необходимости давать в рамках основного доказательства определение «формальной системы» (если такая необходимость все же есть, то см. § 2.7). Вместо этого я воспользуюсь фундаментальным вкладом Тьюринга, который приблизительно в 1936 году описал класс процессов, которые мы сейчас называем «вычислениями» или «алгоритмами» (аналогичные результаты были получены независимо от Тьюринга некоторыми другими математиками, среди которых следует, в первую очередь, упомянуть Черча и Поста). Такие процессы эффективно эквивалентны процедурам, реализуемым в рамках любой математической формальной системы, поэтому для нас не имеет особого значения, что именно понимается под термином «формальная система», коль скоро мы обладаем достаточно ясным представлением о том, что обозначают термины «вычисление» или «алгоритм». Впрочем и для составления такого представления математически строгое определение нам не понадобится.

Те из вас, кто читал мою предыдущую книгу «*Новый разум короля*» (см. НРК, глава 2), возможно, припомнят, что алгоритм там определяется как процедура, которую способна выполнить *машина Тьюринга*, или, если угодно, математически идеализированная вычислительная машина. Такая машина функционирует в пошаговом режиме, причем каждый ее шаг полностью задается нанесенной на рабочую «ленту» меткой, которую (метку) машина «считывает» в соответствующий момент времени, и «внутренним состоянием» машины (дискретно определенным) на этот момент. Количество различных разрешенных внутренних состояний конечно, общее число меток на ленте также должно быть конечным, хотя сама лента по длине не ограничена. Машина начинает работу с какого-то определенного состояния, которое мы обозначим, например, нулем «0», команды же подаются на ленте в виде, скажем, двоичного числа (т. е. последовательности нулей «0» и единиц «1»). Далее машина начинает считывать эти команды, передвигая ленту (либо, что то же самое, перемещаясь вдоль ленты) некоторым определенным образом, согласно встроенным пошаговым инструкциям, при этом действие машины на каждом этапе работы определяется ее внутренним состоянием и конкретным символом, считываемым на данном этапе с ленты. Руководствуясь все теми же встроенными инструкциями, машина может стирать имеющиеся метки или ставить новые. В таком духе машина продолжает работать до тех пор, пока не достигнет особой

команды «STOP», — именно в этот момент (и никак не раньше) машина прекращает работу, а мы можем увидеть на ленте ответ на выполнявшееся вычисление. Вот и все, можно задавать машине новую задачу.

Можно представить себе некую особую машину Тьюринга, которая способна имитировать действие любой возможной машины Тьюринга. Такие машины Тьюринга называют *универсальными*. Иными словами, любая отдельно взятая универсальная машина Тьюринга оказывается в состоянии выполнить *любое* вычисление (или алгоритм), какое нам только может прийти в голову. Хотя внутреннее устройство современного компьютера весьма отличается от устройства описанной выше конструкции (а его внутренняя «рабочая область», пусть и очень велика, все же не бесконечна, в отличие от идеализированной ленты машины Тьюринга), все современные универсальные компьютеры представляют собой, в сущности, универсальные машины Тьюринга.

## 2.2. Вычисления

В этом разделе мы поговорим о *вычислениях*. Под вычислением (или алгоритмом) я подразумеваю действие некоторой машины Тьюринга, или, иными словами, действие компьютера, задаваемое той или иной компьютерной программой. Не следует забывать и о том, что понятие вычисления включает в себя не только выполнение обычных арифметических действий — таких, например, как сложение или умножение чисел, — но и некоторые другие процессы. Так, частью вычислительной процедуры могут стать и вполне определенные *логические операции*. В качестве примера вычисления можно рассмотреть следующую задачу:

**(А)** Найти число, не являющееся суммой квадратов трех чисел.

Под «числом» в данном случае я подразумеваю «натуральное число», т. е. число из ряда

$$0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, \ldots.$$

Под *квадратом* числа понимается результат умножения натурального числа на само себя, т. е. число из ряда

$$0, 1, 4, 9, 16, 25, 36, \ldots;$$



представленные в этом ряду числа получены следующим образом:

$$0 \times 0 = 0^2, \quad 1 \times 1 = 1^2, \quad 2 \times 2 = 2^2, \quad 3 \times 3 = 3^2,$$
$$4 \times 4 = 4^2, \quad 5 \times 5 = 5^2, \quad 6 \times 6 = 6^2, \ldots.$$

Такие числа называются «квадратами», поскольку их можно представить в виде квадратных матриц (пустой матрицей в начале строки обозначен 0):

```
,  *,  * *  ,  * * *  ,  * * * *  ,  ....
   * *     * * *     * * * *
           * * *     * * * *
                     * * * *
```

С учетом вышесказанного решение задачи **(А)** может происходить следующим образом. Мы поочередно проверяем каждое натуральное число, начиная с 0, на предмет того, не является ли оно суммой трех квадратов. При этом, разумеется, рассматриваются только те квадраты, величина которых не превышает самого числа. Таким образом, для каждого натурального числа необходимо проверить некоторое конечное количество квадратов. Отыскав тройку квадратов, составляющих в сумме данное число, переходим к следующему натуральному числу и снова ищем среди квадратов (не превышающих по величине рассматриваемое число) такие три, которые дают в сумме это самое число. Вычисление завершается лишь тогда, когда мы находим натуральное число, которое невозможно получить путем сложения любых трех квадратов. Попробуем применить описанную процедуру на практике и начнем наше вычисление с нуля. Нуль равен $0^2+0^2+0^2$, что, безусловно, является суммой трех квадратов. Далее рассматриваем единицу и находим, что она не равна $0^2 + 0^2 + 0^2$, однако равна $0^2 + 0^2 + 1^2$. Переходим к числу 2 и выясняем, что оно не равно ни $0^2 + 0^2 + 0^2$, ни $0^2 + 0^2 + 1^2$, но равно $0^2+1^2+1^2$. Затем следует число 3 и сумма $3 = 1^2+1^2+1^2$; далее — число 4 и сумма $4 = 0^2 + 0^2 + 2^2$; после $5 = 0^2 + 1^2 + 2^2$ и $6 = 1^2+1^2+2^2$ переходим к 7, и тут обнаруживается, что ни одна из троек квадратов (всех возможных троек квадратов, каждый из которых не превышает 7)

$$0^2+0^2+0^2 \quad 0^2+0^2+1^2 \quad 0^2+0^2+2^2 \quad 0^2+1^2+1^2 \quad 0^2+1^2+2^2$$
$$0^2+2^2+2^2 \quad 1^2+1^2+1^2 \quad 1^2+1^2+2^2 \quad 1^2+2^2+1^2 \quad 2^2+2^2+2^2$$

не дает в сумме 7. На этом этапе вычисление завершается, а мы делаем вывод: 7 есть одно из искомых чисел, так как оно *не* является суммой квадратов трех чисел.

## 2.3. Незавершающиеся вычисления

Будем считать, что с задачей **(A)** нам просто повезло. Попробуем решить еще одну:

**(B)** Найти число, не являющееся суммой квадратов четырех чисел.

На этот раз, добравшись до числа 7, мы находим, что в виде суммы квадратов *четырех* чисел его представить вполне возможно: $7 = 1^2 + 1^2 + 1^2 + 2^2$, поэтому мы переходим к числу 8 (сумма $8 = 0^2 + 0^2 + 2^2 + 2^2$), далее — 9 (сумма $9 = 0^2 + 0^2 + 0^2 + 3^2$) и 10 ($10 = 0^2 + 0^2 + 1^2 + 3^2$) и т. д. Вычисления все продолжаются и продолжаются ($\ldots 23 = 1^2 + 2^2 + 3^2 + 3^2$, $24 = 0^2 + 2^2 + 2^2 + 4^2$, $\ldots$, $359 = 1^2 + 3^2 + 5^2 + 18^2$, $\ldots$) и завершаться, похоже, не собираются. Мы предполагаем, что искомое число, должно быть, невообразимо велико, и для его вычисления нашему компьютеру потребуется чрезвычайно большой промежуток времени и огромный объем памяти. Более того, мы уже начинаем сомневаться, существует ли оно вообще, это самое число. Вычисления все продолжаются и продолжаются, и конца им не видно. Вообще говоря, так оно и есть: описанная вычислительная процедура завершиться в принципе не может. Известна теорема, впервые доказанная в 1770 году великим французским (и отчасти итальянским) математиком Жозефом Луи Лагранжем, согласно которой в виде суммы квадратов четырех чисел можно представить *любое* число. Теорема эта, кстати, весьма непроста (доказать ее как-то пытался великий современник Лагранжа, швейцарский математик Леонард Эйлер, человек, отличавшийся удивительной математической интуицией, оригинальностью и продуктивностью, однако его постигла неудача).

Я, разумеется, не собираюсь докучать читателю подробностями доказательства Лагранжа, вместо этого рассмотрим одну не в пример более простую задачу:

**(C)** Найти нечетное число, являющееся суммой двух четных чисел.



Нисколько не сомневаюсь, что все и так уже все поняли, однако все же поясню. Очевидно, что вычисление, необходимое для решения *этой* задачи, раз начавшись, не завершится никогда. При сложении четных чисел, т. е. чисел, кратных двум,

$$0, 2, 4, 6, 8, 10, 12, 14, 16, \ldots,$$

всегда получаются четные же числа; иными словами, никакая пара четных чисел не может дать в сумме нечетное число, т. е. число вида

$$1, 3, 5, 7, 9, 11, 13, 15, 17, \ldots.$$

Я привел два примера (**(B)** и **(C)**) вычислений, которые невозможно выполнить до конца. Несмотря на то, что в первом случае вычисление и в самом деле никогда не завершается, доказать это довольно непросто, во втором же случае, напротив, бесконечность вычисления более чем очевидна. Позволю себе привести еще один пример:

**(D)** Найти четное число, большее 2, не являющееся суммой двух простых чисел.

Вспомним, что простым называется натуральное число (отличное от 0 и 1), которое делится без остатка лишь само на себя и на единицу; иными словами, простые числа составляют следующий ряд:

$$2, 3, 5, 7, 11, 13, 17, 19, 23, \ldots.$$

Существует довольно высокая вероятность того, что отыскание решения задачи **(D)** также потребует незавершающейся вычислительной процедуры, однако полной уверенности пока нет. Для получения такой уверенности необходимо прежде доказать истинность знаменитой «гипотезы Гольдбаха», выдвинутой Гольдбахом в письме к Эйлеру еще в 1742 году и до сих пор недоказанной.

## 2.4. Как убедиться в невозможности завершить вычисление?

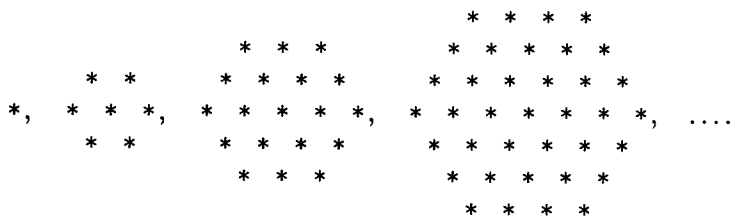
Мы установили, что вычисления могут как успешно завершаться, так и вообще не иметь конца. Более того, в тех случаях, когда вычисление завершиться в принципе не может, это его свойство иногда оказывается очевидным, иногда не совсем

очевидным, а иногда настолько неочевидным, что ни у кого до сих пор не достало сообразительности однозначно такую невозможность доказать. С помощью каких методов математики убеждают самих себя и всех остальных в том, что такое-то вычисление не может завершиться? Применяют ли они при решении подобных задач какие-либо вычислительные (или алгоритмические) процедуры? Прежде чем мы приступим к поиску ответа на этот вопрос, рассмотрим еще один пример. Он несколько менее очевиден, чем **(C)**, но все же гораздо проще **(B)**. Возможно, нам удастся попутно получить некоторое представление о том, с помощью каких средств и методов математики приходят к своим выводам.

В предлагаемом примере участвуют числа, называемые *шестиугольными*:

$$1, 7, 19, 37, 61, 91, 127, \ldots,$$

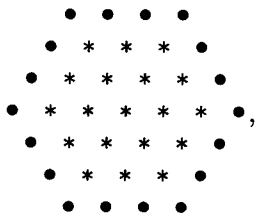
иными словами, числа, из которых можно строить шестиугольные матрицы (пустую матрицу на этот раз мы *не* включаем):



Каждое такое число, за исключением начальной единицы, получается добавлением к предыдущему числу соответствующего числа из ряда кратных 6:

$$6, 12, 18, 24, 30, 36, \ldots.$$

Это легко объяснимо, если обратить внимание на то, что каждое новое шестиугольное число получается путем окружения предыдущего числа шестиугольным кольцом





причем число горошин в этом кольце обязательно будет кратно 6, а множитель при каждом увеличении шестиугольника на одно кольцо будет возрастать ровно на единицу.

Вычислим последовательные *суммы* шестиугольных чисел, увеличивая каждый раз количество слагаемых на единицу, и посмотрим, что из этого получится.

$$1 = 1, \quad 1 + 7 = 8, \quad 1 + 7 + 19 = 27,$$

$$1 + 7 + 19 + 37 = 64, \quad 1 + 7 + 19 + 37 + 61 = 125.$$

Что же особенного в числах 1, 8, 27, 64, 125? Все они являются *кубами*. Кубом называют число, умноженное само на себя трижды:

$$1 = 1^3 = 1 \times 1 \times 1, \quad 8 = 2^3 = 2 \times 2 \times 2, \quad 27 = 3^3 = 3 \times 3 \times 3,$$

$$64 = 4^3 = 4 \times 4 \times 4, \quad 125 = 5^3 = 5 \times 5 \times 5, \ldots.$$

Присуще ли это свойство всем шестиугольным числам? Попробуем следующее число. В самом деле,

$$1 + 7 + 19 + 37 + 61 + 91 = 216 = 6 \times 6 \times 6 = 6^3.$$

Всегда ли выполняется это правило? Если да, то никогда не завершится вычисление, необходимое для решения следующей задачи:

**(E)** Найти последовательную сумму шестиугольных чисел, начиная с единицы, не являющуюся кубом.

Думается, я сумею убедить вас в том, что это вычисление и в самом деле можно выполнять вечно, но так и не получить искомого ответа.

Прежде всего отметим, что число называется кубом не просто так: из соответствующего количества точек можно сложить трехмерный массив в форме куба (такой, например, как на рис. 2.1). Попробуем представить себе построение такого массива в виде последовательности шагов: вначале разместим где-нибудь угловую точку, а затем будем добавлять к ней, одну за другой, особые конфигурации точек, составленные из трех «плоскостей» — задней стенки, боковой стенки и потолка, как показано на рис. 2.2.

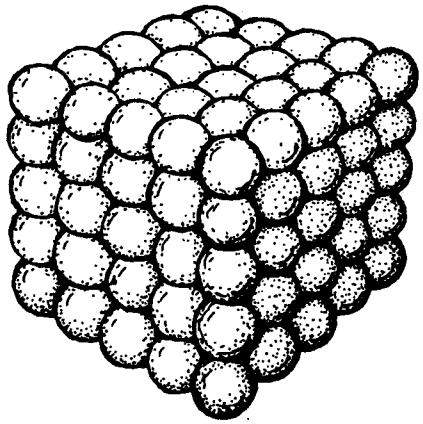


Рис. 2.1. Сферы, уложенные в кубический массив.

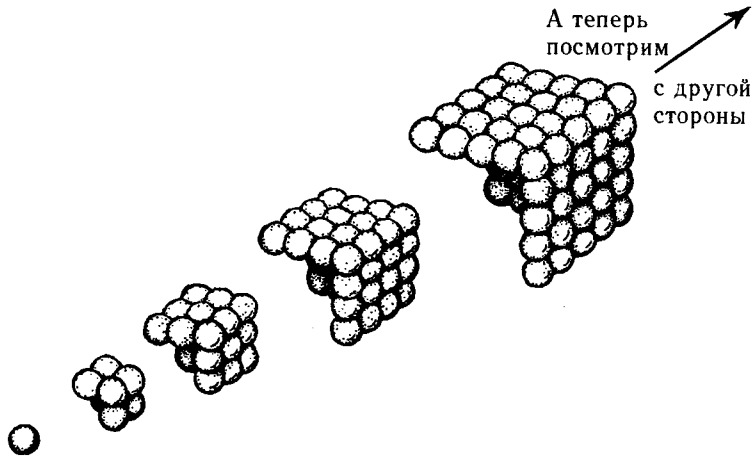



Рис. 2.2. Разберем куб на части — каждая со своей задней стенкой, боковой стенкой и потолком.

Посмотрим теперь на одну из наших трехгранных конфигураций со стороны, т. е. вдоль прямой, соединяющей начальную точку построения и точку, общую для всех трех граней. Мы уви-



дим *шестиугольник*, подобный тому, что изображен на рис. 2.3. Точки, из которых складываются эти увеличивающиеся в размере шестиугольники, представляют собой, в сущности, те же точки, что образуют полный куб. То есть получается, что последовательное сложение шестиугольных чисел, начиная с единицы, всегда будет давать число кубическое. Следовательно, можно считать доказанным, что вычисление, требуемое для решения задачи **(E)**, никогда не завершится.

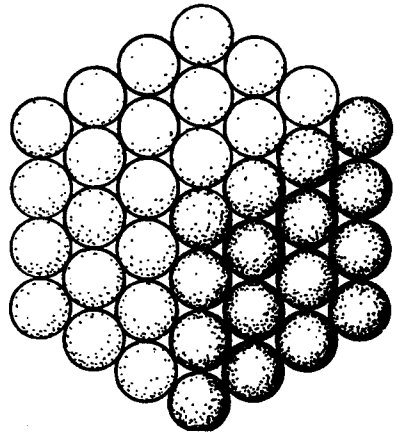


Рис. 2.3. Каждую часть построения можно рассматривать как шестиугольник.

Кто-то, быть может, уже готов упрекнуть меня в том, что представленные выше рассуждения можно счесть в лучшем случае интуитивным умозаключением, но не формальным и строгим математическим доказательством. На самом же деле, перед вами именно доказательство, и доказательство вполне здравое, а пишу все это я отчасти и для того, чтобы показать, что осмысленность того или иного метода математического обоснования никак не связана с его «формализованностью» в соответствии с какой-либо заранее заданной и общепринятой системой правил. Напомню, кстати, о еще более элементарном примере геометрического обоснования, применяемого для получения одного общего свойства натуральных чисел, — речь идет о доказательстве истинно-

сти равенства $a \times b = b \times a$, приведенном в § 1.19. Тоже вполне достойное «доказательство», хотя формальным его назвать нельзя.

Представленное выше рассуждение о суммировании последовательных шестиугольных чисел можно при желании заменить более формальным математическим доказательством. В основу такого формального доказательства можно положить *принцип математической индукции*, т. е. процедуру установления истинности утверждения в отношении *всех* натуральных чисел на основании одного-единственного вычисления. По существу, этот принцип позволяет заключить, что некое положение $P(n)$, зависящее от конкретного натурального числа $n$ (например, такое: «сумма первых $n$ шестиугольных чисел равна $n^3$»), справедливо для *всех* $n$, если мы можем показать, во-первых, что оно справедливо для $n = 0$ (или, в нашем случае, для $n = 1$), и, во-вторых, что из истинности $P(n)$ *следует* истинность и $P(n+1)$. Думаю, нет необходимости описывать здесь в деталях, как можно с помощью математической индукции доказать невозможность завершить вычисление **(E)**; тем же, кого данная тема заинтересовала, рекомендую попытаться в качестве упражнения выполнить такое доказательство самостоятельно.

Всегда ли для установления факта действительной незавершаемости вычисления достаточно применить некие четко определенные правила — такие, например, как принцип математической индукции? Как ни странно, нет. Это утверждение, как мы вскоре увидим, является одним из следствий теоремы Гёделя, и для нас крайне важно попытаться его правильно понять. Причем недостаточной оказывается не только математическая индукция. Недостаточным будет *какой угодно* набор правил, если под «набором правил» подразумевать некую систему формализованных процедур, в рамках которой возможно исключительно вычислительным путем проверить корректность применения этих правил в каждом конкретном случае. Такой вывод может показаться чересчур пессимистичным, ибо он, по-видимому, означает, что, несмотря на то, что вычисления, которые нельзя завершить, существуют, сам факт их незавершаемости строго математически установить невозможно. Однако смысл упомянутого следствия из теоремы Гёделя заключается вовсе не в этом. *На самом деле*, все не так уж и плохо: способность понимать и делать выводы, присущая математикам — как, впрочем, и всем остальным людям, наделенным логическим мышлением и воображением, — просто-



напросто не поддается формализации в виде того или иного набора правил. Иногда правила могут стать частичной заменой пониманию, однако в полной мере такая замена не представляется возможной.

## 2.5. Семейства вычислений; следствие Гёделя – Тьюринга $\mathscr{G}$

Для того, чтобы понять, каким образом из теоремы Гёделя (в моей упрощенной формулировке, навеянной отчасти идеями Тьюринга) следует все вышесказанное, нам необходимо будет сделать небольшое обобщение для типов утверждений, относящихся к рассмотренным в предыдущем разделе вычислениям. Вместо того чтобы решать проблему завершаемости для каждого отдельного вычисления (**(A)**, **(B)**, **(C)**, **(D)** или **(E)**), нам следует рассмотреть некоторое общее вычисление, которое зависит от *натурального числа* $n$ (либо как-то *воздействует* на него). Таким образом, обозначив такое вычисление через $C(n)$, мы можем рассматривать его как целое *семейство* вычислений, где для каждого натурального числа (0, 1, 2, 3, 4, ...) выполняется отдельное вычисление (соответственно, $C(0)$, $C(1)$, $C(2)$, $C(3)$, $C(4)$, ...), а сам принцип, в соответствии с которым вычисление зависит от $n$, является целиком и полностью вычислительным.

В терминах машин Тьюринга это всего лишь означает, что $C(n)$ есть действие, производимое некоей машиной Тьюринга над числом $n$. Иными словами, число $n$ наносится на ленту и подается на вход машины, после чего машина самостоятельно выполняет вычисления. Если вас почему-либо не устраивает концепция «машины Тьюринга», вообразите себе самый обыкновенный универсальный компьютер и считайте $n$ «данными», необходимыми для работы какой-нибудь программы. Нас в данном случае интересует лишь одно: при любом ли значении $n$ может завершиться работа такого компьютера.

Для того чтобы пояснить, что именно понимается под вычислением, зависящим от натурального числа $n$, рассмотрим два примера:

**(F)** найти число, не являющееся суммой квадратов $n$ чисел,

и

**(G)** найти нечетное число, являющееся суммой $n$ четных чисел.

Припомнив, о чем говорилось выше, мы без особого труда убедимся, что вычисление **(F)** завершается *только* при $n = 0, 1, 2$ и 3 (давая в результате, соответственно, 1, 2, 3 и 7), тогда как вычисление **(G)** вообще не завершается ни при каком значении $n$. Вздумай мы действительно доказать, что вычисление **(F)** не завершается при $n$, равном или большем 4, нам понадобилась бы более или менее серьезная математическая подготовка (по крайней мере, знакомство с доказательством Лагранжа); с другой стороны, тот факт, что ни при каком $n$ не завершается вычисление **(G)**, вполне очевиден. Какими же процедурами располагают математики для установления незавершаемой природы таких вычислений в общем случае? Можно ли сами эти процедуры представить в вычислительной форме?

Предположим, что у нас имеется некая вычислительная процедура $A$, которая по завершении[1] дает нам исчерпывающее доказательство того, что вычисление $C(n)$ действительно никогда не заканчивается. Ниже мы попробуем вообразить, что $A$ включает в себя *все* известные математикам процедуры, посредством которых можно убедительно доказать, что то или иное вычисление никогда не завершается. Соответственно, если в каком-то конкретном случае завершается процедура $A$, то мы получаем, в рамках доступного человеку знания, доказательство того, что рассматриваемое конкретное вычисление никогда *не* заканчивается. Большая часть последующих рассуждений не потребует участия процедуры $A$ именно в такой роли, так как они посвящены, в основном, математическим умопостроениям. Однако для получения окончательного заключения $\mathscr{G}$ нам придется-таки придать процедуре $A$ соответствующий статус.

Я, разумеется, не требую, чтобы посредством процедуры $A$ всегда можно было однозначно установить, что вычисление $C(n)$ нельзя завершить (в случае, если это действительно так); однако я настаиваю на том, что неверных ответов $A$ не дает, т. е. если мы с ее помощью пришли к выводу, что вычисление $C(n)$ не завершается, значит, так оно и есть. Процедуру $A$, которая и в самом деле всегда дает верный ответ, мы будем называть *обоснованной*.

[1] Здесь я предполагаю, что если процедура $A$ вообще завершается, то это свидетельствует об успешном установлении факта незавершаемости $C(n)$. Если же $A$ «застревает» по какой-либо иной, нежели достижение «успеха», причине, то это означает, что в данном случае процедура $A$ корректно завершиться не может. См. далее по тексту возражения **Q3** и **Q4**, а также Приложение А, с. 193.



Следует отметить, что если процедура $A$ оказывается в действительности необоснованной, то этот факт, в принципе, можно установить с помощью прямого вычисления — иными словами, необоснованную процедуру $A$ можно опровергнуть вычислительными методами: если $A$ ошибочно утверждает, что вычисление $C(n)$ нельзя завершить, тогда как в действительности это не так, то выполнение самого вычисления $C(n)$ в конечном счете приведет к опровержению $A$. (Возможность практического выполнения такого вычисления представляет собой отдельный вопрос, его мы рассмотрим в ответе на возражение **Q8**.)

Для того чтобы процедуру $A$ можно было применять к вычислениям в общем случае, нам потребуется какой-нибудь способ маркировки различных вычислений $C(n)$, допускаемый $A$. Все возможные вычисления $C$ можно, вообще говоря, представить в виде простой последовательности

$$C_0, C_1, C_2, C_3, C_4, C_5, \ldots,$$

т. е. $q$-е вычисление при этом получит обозначение $C_q$. В случае применения такого вычисления к конкретному числу $n$ будем записывать

$$C_0(n), C_1(n), C_2(n), C_3(n), C_4(n), C_5(n), \ldots.$$

Можно представить, что эта последовательность задается, скажем, как некий пронумерованный ряд компьютерных программ. (Для большей ясности мы могли бы, при желании, рассматривать такую последовательность как ряд пронумерованных машин Тьюринга, описанных в НРК; в этом случае вычисление $C_q(n)$ представляет собой процедуру, выполняемую $q$-й машиной Тьюринга $T_q$ над числом $n$.) Здесь важно учитывать следующий технический момент: рассматриваемая последовательность является *вычислимой* — иными словами, существует одно-единственное[2] вычисление $C_\bullet$, которое, будучи выполнено над числом $q$, дает в результате $C_q$, или, если точнее, выполнение вычисления $C_\bullet$ над *парой* чисел $q$, $n$ (именно в таком порядке) дает в результате $C_q(n)$.

[2] Собственно, точно такой же результат достигается посредством процедуры, выполняемой универсальной машиной Тьюринга над парой чисел $q$, $n$; см. Приложение А и НРК, с. 51–57.

Можно полагать, что процедура $A$ представляет собой некое особое вычисление, выполняя которое над парой чисел $q$, $n$, можно однозначно установить, что вычисление $C_q(n)$, в конечном итоге, никогда не завершится. Таким образом, когда *завершается* вычисление $A$, мы имеем достаточное доказательство того, что вычисление $C_q(n)$ *завершить невозможно*. Хотя, как уже говорилось, мы и попытаемся вскоре представить себе такую процедуру $A$, которая формализует *все* известные современной математике процедуры, способные достоверно установить невозможность завершения вычисления, нет никакой необходимости придавать $A$ такой смысл прямо сейчас. Пока же процедурой $A$ мы будем называть *любой обоснованный* набор вычислительных правил, с помощью которого можно установить, что то или иное вычисление $C_q(n)$ никогда не завершается. Поскольку выполняемое процедурой $A$ вычисление зависит от двух чисел $q$ и $n$, его можно обозначить как $A(q, n)$ и записать следующее утверждение:

**(H)** Если завершается $A(q, n)$, то $C_q(n)$ не завершается.

Рассмотрим частный случай утверждения **(H)**, положив $q$ равным $n$. Такой шаг может показаться странным, однако он вполне допустим. (Он представляет собой первый этап мощного «диагонального доказательства» — процедуры, открытой в высшей степени оригинальным и влиятельным датско-русско-немецким математиком девятнадцатого века Георгом Кантором; эта процедура лежит в основе рассуждений и Гёделя, и Тьюринга.) При $q$, равном $n$, наше утверждение принимает следующий вид:

**(I)** Если завершается $A(n, n)$, то $C_n(n)$ не завершается.

Отметим, что $A(n, n)$ зависит только от *одного* числа $(n)$, а не от двух, так что данное вычисление должно принадлежать ряду $C_0, C_1, C_2, C_3, \ldots$ (по $n$), поскольку предполагается, что этот ряд содержит *все* вычисления, которые можно выполнить над одним натуральным числом $n$. Обозначив это вычисление через $C_k$, запишем:

**(J)** $A(n, n) = C_k(n)$.

Рассмотрим теперь частный случай $n = k$. (Второй этап диагонального доказательства Кантора.) Из равенства **(J)** получаем:

**(K)** $A(k, k) = C_k(k)$,



утверждение же **(I)** при $n = k$ принимает вид:

**(L)** Если завершается $A(k, k)$, то $C_k(k)$ не завершается.

Подставляя **(K)** в **(L)**, находим:

**(M)** Если завершается $C_k(k)$, то $C_k(k)$ не завершается.

Из этого следует заключить, что вычисление $C_k(k)$ в действительности *не* завершается. (Ибо, согласно **(M)**, если оно завершается, то оно не завершается!) Невозможно завершить и вычисление $A(k, k)$, поскольку, согласно **(K)**, оно *совпадает* с $C_k(k)$. То есть наша процедура $A$ оказывается не в состоянии показать, что данное конкретное вычисление $C_k(k)$ не завершается, даже если оно и в самом деле не завершается.

Более того, если нам *известно*, что процедура $A$ обоснованна, то, значит, нам *известно* и то, что вычисление $C_k(k)$ не завершается. Иными словами, нам известно нечто, о чем посредством процедуры $A$ мы узнать не могли. Следовательно, сама процедура $A$ с нашим пониманием *никак* не связана. ■

В этом месте осторожный читатель, возможно, пожелает перечесть все вышеприведенное доказательство заново, дабы убедиться в том, что он не пропустил какой-нибудь «ловкости рук» с моей стороны. Надо признать, что, на первый взгляд, это доказательство и в самом деле смахивает на фокус, и все же оно полностью допустимо, а при более тщательном изучении лишь выигрывает в убедительности. Мы обнаружили некое вычисление $C_k(k)$, которое, насколько нам известно, не завершается; однако установить этот факт с помощью имеющейся в нашем распоряжении вычислительной процедуры $A$ мы не в состоянии. Это, собственно, и есть теорема Гёделя(—Тьюринга) в необходимом мне виде. Она применима к любой вычислительной процедуре $A$, предназначенной для установления невозможности завершить вычисление, — *коль скоро нам известно, что упомянутая процедура обоснованна*. Можно заключить, что для однозначного установления факта незавершаемости вычисления не будет вполне достаточным ни один из заведомо обоснованных наборов вычислительных правил (такой, например, как процедура $A$), поскольку существуют незавершающиеся вычисления (например, $C_k(k)$), на которые эти правила не распространяются. Более того, поскольку на основании того, что нам известно о процедуре $A$ и об ее обоснованности, мы действительно можем

составить вычисление $C_k(k)$, которое, *очевидно*, никогда не завершается, мы вправе заключить, что процедуру $A$ никоим образом *нельзя* считать формализацией процедур, которыми располагают математики для установления факта незавершаемости вычисления, вне зависимости от конкретной природы $A$. Вывод:

$\mathscr{G}$ Для установления математической истины математики не применяют заведомо обоснованные алгоритмы.

Мне представляется, что к такому выводу неизбежно должен прийти всякий логически рассуждающий человек. Однако многие до сих пор предпринимают попытки этот вывод опровергнуть (выдвигая возражения, обобщенные мною под номерами **Q1**–**Q20** в § 2.6 и § 2.10), и, разумеется, найдется ничуть не меньше желающих оспорить вывод более строгий, суть которого сводится к тому, что мыслительная деятельность непременно оказывается связана с некими феноменами, носящими фундаментально невычислительный характер. Вы, возможно, уже спрашиваете себя, каким же это образом подобные математические рассуждения об абстрактной природе вычислений могут способствовать объяснению принципов функционирования человеческого мозга. Какое такое отношение имеет все вышесказанное к проблеме осмысленного осознания? Дело в том, что, благодаря этим математическим рассуждениям, мы и впрямь можем прояснить для себя некие весьма важные аспекты такого свойства мышления, как *понимание* — в терминах общей вычислимости, — а как было показано в § 1.12, свойство понимания связано с осмысленным осознанием самым непосредственным образом. Предшествующее рассуждение действительно носит в основном математический характер, и связано это с необходимостью подчеркнуть одно очень существенное обстоятельство: алгоритм $A$ участвует здесь на двух совершенно различных уровнях. С одной стороны, это просто некий алгоритм, обладающий определенными свойствами; с другой стороны, получается, что *на самом-то деле* $A$ можно рассматривать как «алгоритм, которым пользуемся мы сами» в процессе установления факта незавершаемости того или иного вычисления. Так что в вышеприведенном рассуждении речь идет не только и не столько о вычислениях. Речь идет также и о том, каким образом мы используем нашу способность к осмысленному пониманию для составления заключения об истинности какого-либо математического утверждения — в данном случае утверждения о незавершаемости вычисления $C_k(k)$.



Именно взаимодействие между двумя различными уровнями рассмотрения алгоритма $A$ — в качестве гипотетического способа функционирования сознания и собственно вычисления — позволяет нам сделать вывод, выражающий фундаментальное противоречие между такой сознательной деятельностью и простым вычислением.

Существуют, однако, всевозможные лазейки и контраргументы, на которые необходимо обратить самое пристальное внимание. Для начала, в оставшейся части этой главы, я тщательно разберу *все* важные контраргументы против вывода $\mathscr{G}$, которые когда-либо попадались мне на глаза — см. возражения **Q1**–**Q20** и комментарии к ним в §§ 2.6 и 2.10; там, кроме того, можно найти и несколько дополнительных возражений моего собственного изобретения. Каждое из возражений будет разобрано со всей обстоятельностью, на какую я только способен. Пройдя через это испытание, вывод $\mathscr{G}$, как мы убедимся, существенно не пострадает. Далее, в главе 3, я рассмотрю следствия уже из утверждения $\mathscr{G}$. Мы обнаружим, что оно и в самом деле способно послужить прочным фундаментом для построения весьма убедительного доказательства *абсолютной* невозможности точного моделирования сознательного математического понимания посредством вычислительных процедур, будь то восходящие, нисходящие или любые их сочетания. Многие сочтут такой вывод весьма неприятным, поскольку если он справедлив, то нам, получается, просто некуда двигаться дальше. Во второй части книги я выберу более позитивный курс. Я приведу правдоподобные, на мой взгляд, научные доводы в пользу справедливости результатов моих размышлений о физических процессах, которые могут, предположительно, лежать в основе деятельности мозга — вроде той, что осуществляется при нашем восприятии приведенных выше рассуждений, — и о причинах недоступности этой деятельности для какого бы то ни было вычислительного описания.

## 2.6. Возможные формальные возражения против $\mathscr{G}$

Утверждение $\mathscr{G}$ вполне способно потрясти воображение и не слишком впечатлительного читателя, особенно если учесть достаточно простой характер составных элементов рассуждения, из

которого мы это утверждение вывели. Прежде чем перейти к рассмотрению (в главе 3) его следствий применительно к возможности создания разумного робота-математика с компьютерным разумом, необходимо очень тщательно исследовать некоторое количество формальных моментов, связанных с получением вывода $\mathscr{G}$. Если подобные возможные формальные «лазейки» вас не смущают и вы готовы принять на веру утверждение $\mathscr{G}$ (согласно которому, напомним, математики при установлении математической истины не применяют заведомо обоснованные алгоритмы), то вы, вероятно, предпочтете пропустить (или хотя бы на некоторое время отложить) нижеследующие рассуждения и перейти непосредственно к главе 3. Более того, если вы готовы принять на веру и несколько более серьезный вывод, в соответствии с которым *принципиально* невозможно алгоритмически объяснить ни математическое, ни какое-либо иное понимание, то вам, возможно, стоит перейти сразу ко второй части книги — задержавшись разве что на воображаемом диалоге в § 3.23 (обобщающем наиболее важные аргументы главы 3) и выводах в § 3.28.

Существует несколько математических моментов, связанных с приведенным в § 2.5 гёделевским доказательством, которые не дают людям покоя. Попытаемся с этими моментами разобраться.

**Q1. Я понимаю так, что процедура $A$ является *единичной*, тогда как во всевозможных математических обоснованиях мы, несомненно, применяем много разных способов рассуждения. Не следует ли нам принять во внимание возможность существования целого ряда возможных «процедур $A$»?**

В действительности, использование мною такой формулировки вовсе не влечет за собой потери общего характера рассуждений в целом. Любой конечный ряд $A_1$, $A_2$, $A_3$, ..., $A_r$ алгоритмических процедур всегда можно выразить в виде единичного алгоритма $A$, причем таким образом, что $A$ окажется незавершаемым только в том случае, если не завершаются *все* отдельные алгоритмы $A_1$, ..., $A_r$. (Процедура $A$ может протекать, например, следующим образом: «Выполнить первые 10 шагов алгоритма $A_1$; запомнить результат; выполнить первые 10 шагов алгоритма $A_2$; запомнить результат; выполнить первые 10 шагов алгоритма $A_3$; запомнить результат; и так далее вплоть до $A_r$; затем вернуться к $A_1$ и выполнить следующие 10 шагов; запомнить результат и т. д.; затем перейти к третьей группе из 10 шагов и т. п. Завершить процедуру, как только завершится любой из алгоритмов $A_r$».) Если же ряд алгоритмов $A$ бесконечен, то для того, чтобы его можно было считать алгоритмической процедурой, необходимо найти способ порождения всей совокупности алгоритмов $A_1$, $A_2$, $A_3$, ... алгоритмическим путем. Тогда мы сможем получить единичный алгоритм $A$, который заменяет весь ряд алгоритмов и выглядит приблизительно следующим образом:

«первые 10 этапов $A_1$;

вторые 10 этапов $A_1$, первые 10 этапов $A_2$;

третьи 10 этапов $A_1$, вторые 10 этапов $A_2$, первые 10 этапов $A_3$;

... и т. д.» ....

Завершается такой алгоритм лишь после успешного завершения любого алгоритма из ряда, и никак не раньше.

С другой стороны, можно представить себе ситуацию, когда ряд $A_1$, $A_2$, $A_3$, ..., предположительно бесконечный, заранее не задан даже в принципе. Время от времени к такому ряду добавляется следующая алгоритмическая процедура, однако изначально весь ряд в целом не определен. В этом случае, ввиду отсутствия какой-либо предварительно заданной алгоритмической процедуры для порождения такого ряда, единичный замкнутый алгоритм нам получить никак не удастся.

**Q2. Мы, безусловно, должны допустить, что алгоритм $A$ может оказаться и не фиксированным. Люди, в конце концов, обладают способностью к обучению, а значит, применяемый ими при этом алгоритм вполне может претерпевать непрерывные изменения.**

Для описания изменяющегося алгоритма необходимо каким-то образом задать правила, согласно которым он, собственно, изменяется. Если сами по себе эти правила являются полностью алгоритмическими, то мы уже включили их в описание нашей гипотетической процедуры «$A$», иначе говоря, *такой* «изменяющийся алгоритм» на деле представляет собой всего-навсего

еще один пример единичного алгоритма, и на наши рассуждения подобное допущение никак не влияет. С другой стороны, можно вообразить средства для изменения алгоритма, предположительно *не* являющиеся алгоритмическими: такие, например, как введение в алгоритм каких-то случайных составляющих или неких процедур взаимодействия его с окружением. «Неалгоритмический» статус подобных средств изменения алгоритма мы еще будем рассматривать несколько позднее (см. §§ 3.9, 3.10); можно также вернуться к § 1.9, где было показано, что ни одно из этих средств не позволяет сколько-нибудь убедительно избавиться от алгоритмизма[3] (как того требует точка зрения $\mathscr{C}$). В данном случае, т. е. в рамках чисто математических рассуждений, нас занимает лишь возможность того, что такое изменение действительно будет носить алгоритмический характер. Если же предположить, что алгоритмическим оно быть никак *не может*, то мы, безусловно, придем к полному согласию с выводом $\mathscr{G}$.

Пожалуй, следует немного подробнее остановиться на том, что может обозначать определение «алгоритмически изменяющийся» применительно к алгоритму $A$. Допустим, что алгоритм $A$ зависит не только от $q$ и $n$, но и еще от одного параметра $t$, который можно рассматривать как «время», а можно как просто количество предшествующих настоящему моменту случаев активации нашего алгоритма. Как бы то ни было, мы можем также предположить, что параметр $t$ является натуральным числом, и записать следующий ряд алгоритмов $A_t(q, n)$:

$$A_0(q, n), \quad A_1(q, n), \quad A_2(q, n), \quad A_3(q, n), \quad \ldots,$$

каждый элемент которого предположительно является обоснованной процедурой для установления незавершаемости вычисления $C_q(n)$; при этом мы будем считать, что мощность этих процедур возрастает по мере увеличения $t$. Предполагается также, что способ, посредством которого увеличивается мощность этих процедур, является алгоритмическим. Возможно, этот «алгоритмический способ» зависит некоторым образом от «опыта» выполнения предыдущих алгоритмов $A_t(q, n)$, однако в данном случае мы предполагаем, что этот «опыт» порождается также алгоритмически (в противном случае мы снова приходим к согласию с $\mathscr{G}$),

[3]Термин «алгоритмизм», который (по своей сути) прекрасно подходит для обозначения «точки зрения $\mathscr{A}$» в моей классификации, был предложен Хао Ваном [377].



т. е. мы имеем полное право включить «опыт» (или способы его порождения) в перечень операций, составляющих следующий алгоритм (т. е., собственно, в $A_t(q, n)$). Действуя таким образом, мы опять-таки получаем *единичный* алгоритм ($A_t(q, n)$), который зависит алгоритмически от всех *трех* параметров: $t$, $q$, $n$. На его основе можно построить алгоритм $A^*$, столь же мощный, что и весь ряд $A_t(q, n)$, однако зависящий только от двух натуральных чисел: $q$ и $n$. Для получения такого $A^*(q, n)$ нам, как и прежде, необходимо лишь выполнить первые десять шагов алгоритма $A_0(q, n)$ и запомнить результат; затем первые десять шагов алгоритма $A_1(q, n)$ и вторые десять шагов алгоритма $A_0(q, n)$, запоминая получаемые результаты; затем первые десять шагов алгоритма $A_2(q, n)$, вторые десять шагов алгоритма $A_1(q, n)$, третьи десять шагов алгоритма $A_0(q, n)$ и т. д., запоминая получаемые на каждом шаге вычисления результаты. В конечном итоге, сразу после завершения *любого* из составляющих алгоритм вычислений завершается выполнение и *всей процедуры в целом*. Замена процедуры $A$ процедурой $A^*$ никак не влияет на ход рассуждений, посредством которых мы пришли к выводу $\mathscr{G}$.

**Q3. Не был ли я излишне категоричен, утверждая, что в тех случаях, когда уже можно определенно утверждать, что данное вычисление $C_q(n)$ и вправду *завершается*, алгоритм $A$ все равно должен выполняться бесконечно? Допусти мы, что $A$ в таких случаях также *завершается*, все наше рассуждение оказалось бы ложным. В конце концов, общеизвестно, что присущая людям способность к интуитивному пониманию позволяет им порой делать заключение о возможности завершения того или иного вычисления, однако я, судя по всему, здесь этой способностью пренебрег. Не слишком ли много искусственных ограничений?**

Вовсе нет. Предполагается, что наше рассуждение применимо лишь к тому пониманию, которое позволяет заключить, что вычисление *не* завершается, но никак не к тому пониманию, благодаря которому мы приходим к противоположному выводу. Гипотетический алгоритм $A$ вовсе не обязан достигать «успешного завершения», обнаружив что то или иное вычисление *завершается*. Не в этом заключается его смысл.

Если вас такое положение дел не устраивает, попробуйте представить алгоритм $A$ следующим образом: пусть $A$ объединяет в себе *оба* вида понимания, но в том случае, когда выясняется, что вычисление $C_q(n)$ действительно завершается, алгоритм $A$ искусственно зацикливается (т. е. выполняет какую-то операцию снова и снова, бесконечное количество раз). Разумеется, на самом деле математики работают иначе, однако дело не в этом. Наше рассуждение построено как *reductio ad absurdum*[4], т. е. начав с допущения, что для установления математической истины используются заведомо обоснованные алгоритмы, мы в итоге приходим к противоположному выводу. Такое доказательство не требует, чтобы гипотетическим алгоритмом непременно оказался какой-то конкретный алгоритм $A$, мы вполне можем заменить его на другой алгоритм, построенный на основе $A$, — как, например, в только что упомянутом случае.

Этот комментарий применим и к любому другому возражению вида: «А что если алгоритм $A$ завершится по какой-либо совершенно посторонней причине и не даст нам доказательства того, что вычисление $C_q(n)$ не завершается?». Если нам вдруг придется иметь дело с алгоритмом «$A$», который ведет себя подобным образом, то мы просто применим представленное в §2.5 обоснование к немного другому $A$ — к такому, который зацикливается всякий раз, когда исходный «$A$» завершается по любой из упомянутых посторонних причин.

**Q4. Судя по всему, каждое вычисление $C_q$ в предложенной мною последовательности $C_0, C_1, C_2, \ldots$ является вполне определенным, тогда как при любом прямом переборе (численном или алфавитном) компьютерных программ ситуация, конечно же, была бы иной?**

В самом деле, было бы весьма затруднительно однозначно гарантировать, что каждому натуральному числу $q$ в нашей последовательности действительно соответствует некое рабочее вычисление $C_q$. Например, описанная в НРК последовательность машин Тьюринга $T_q$ этому условию, конечно же, не удовлетворяет; см. НРК, с. 54. При определенных значениях $q$ машину Тьюринга $T_q$ можно назвать «фиктивной» по одной из четырех причин: ее работа никогда не завершается; она оказывается «некорректно определенной», поскольку представление числа $n$ в виде двоичной последовательности содержит слишком много (пять или более) единиц подряд и, как следствие, не имеет интерпретации в данной схеме; она получает команду, которая вводит ее в нигде не описанное внутреннее состояние; или же по завершении работы она оставляет ленту пустой, т. е. не дает никакого численно интерпретируемого результата. (См. также Приложение А.) Для приведенного в §2.5 доказательства Гёделя–Тьюринга вполне достаточно объединить все эти причины в одну категорию под названием «вычисление не завершается». В частности, когда я говорю, что вычислительная процедура $A$ «завершается» (см. также примечание на с. 124), я подразумеваю, что она «завершается» как раз в вышеупомянутом смысле (а потому не содержит неинтерпретируемых последовательностей и не оставляет ленту пустой), — иными словами, «завершиться» может только действительно корректно определенное рабочее вычисление. Аналогично, фраза «вычисление $C_q(n)$ завершается» означает, что данное вычисление корректно завершается именно в этом смысле. При такой интерпретации соображение **Q4** не имеет совершенно никакого отношения к представленному мною доказательству.

[4] Приведение к абсурду (лат.), доказательство от противного. — *Прим. перев.*



**Q5. Не является ли мое рассуждение лишь демонстрацией неприменимости некоей *частной* алгоритмической процедуры ($A$) к выполнению вычисления $C_q(n)$? И каким образом оно показывает, что я справлюсь с задачей лучше, чем какая бы то ни было процедура $A$?**

Оно *и в самом деле* вполне однозначно показывает, что мы справляемся с такого рода задачами гораздо лучше *любого* алгоритма. Поэтому, собственно, я и воспользовался в своем рассуждении приемом *reductio ad absurdum*. Пожалуй, в данном случае уместно будет привести аналогию. Читателям, вероятно, известно о евклидовом доказательстве невозможности отыскать наибольшее простое число, также основанном на *reductio ad absurdum*. Доказательство Евклида выглядит следующим образом. Допустим обратное: такое наибольшее простое число нам известно; назовем его $p$. Теперь рассмотрим число $N$, которое представляет собой сумму произведения всех простых чисел

вплоть до $p$ и единицы:

$$N = 2 \times 3 \times 5 \times \ldots \times p + 1.$$

Число $N$, безусловно, больше $p$, однако оно не делится ни на одно из простых чисел 2, 3, 5, ..., $p$ (поскольку при делении получаем единицу в остатке), откуда следует, что $N$ либо и есть искомое наибольшее простое число, либо оно является составным, и тогда его можно разделить на простое число, большее $p$. И в том, и в другом случае мы находим простое число, большее $p$, что противоречит исходному допущению, заключавшемуся в том, что $p$ есть наибольшее простое число. Следовательно, наибольшее простое число отыскать нельзя.

Такое рассуждение, основываясь на *reductio ad absurdum*, не просто показывает, что требуемому условию не соответствует некое *частное* простое число $p$, поскольку можно отыскать число больше него; оно показывает, что наибольшего простого числа просто *не может* существовать в природе. Аналогично, представленное выше доказательство Гёделя–Тьюринга не просто показывает, что нам не подходит тот или иной *частный* алгоритм $A$, оно демонстрирует, что *в природе не существует* алгоритма (познаваемо обоснованного), который был бы эквивалентен способности человека к интуитивному пониманию, которую мы применяем для установления факта незавершаемости тех или иных вычислений.

**Q6. Можно составить программу, выполняя которую, компьютер в точности повторит все этапы представленного мною доказательства. Не означает ли это, что компьютер оказывается в состоянии самостоятельно прийти к любому заключению, к какому пришел бы я сам?**

Отыскание конкретного вычисления $C_k(k)$ при заданном алгоритме $A$, безусловно, представляет собой вычислительный процесс. Более того, это можно достаточно явно показать[5]. Означает ли это, что предположительно неалгоритмическая математическая интуиция — интуиция, благодаря которой мы определяем, что вычисление $C_k(k)$ никогда не завершается, — на деле является все же алгоритмической?

[5] Чтобы подчеркнуть, что я принимаю это обстоятельство во внимание, я отсылаю читателя к Приложению А, где представлена явная вычислительная процедура (выполненная в соответствии с правилами, подробно описанными в НРК, глава 2) для получения операции $C_k(k)$ машины Тьюринга посредством алгоритма $A$. Здесь предполагается, что алгоритм $A$ задан в виде машины Тьюринга $T_a$, определение же вычисления $C_q(n)$ кодируется как операция машины $T_a$ над числом $q$, а затем над числом $n$.

Плохой перевод? Здесь и далее.



Думаю, данное суждение следует рассмотреть более подробно, поскольку оно представляет собой одно из наиболее распространенных недоразумений, связанных с гёделевским доказательством. Следует особо уяснить, что оно *не сводит на нет* ничего из сказанного ранее. Хотя процедуру отыскания вычисления $C_k(k)$ с помощью алгоритма $A$ можно представить в виде вычисления, это вычисление не входит в перечень процедур, содержащихся в $A$. И не может входить, поскольку самостоятельно алгоритм $A$ не способен установить истинность $C_k(k)$, тогда как новое вычисление (вкупе с $A$), судя по всему, вполне на это способно. Таким образом, несмотря на то, что с помощью нового вычисления действительно можно отыскать вычисление $C_k(k)$, членом клуба «официальных установителей истины» оно не является.

Изложим все это несколько иначе. Вообразите себе управляемого компьютером робота, способного устанавливать математические истины с помощью алгоритмических процедур, содержащихся в $A$. Для большей наглядности я буду пользоваться антропоморфной терминологией и говорить, что робот «знает» те математические истины (в данном случае — связанные с установлением факта незавершаемости вычислений), которые он может вывести, применяя алгоритм $A$. Однако если наш робот «знает» лишь $A$, то он *никак не сможет* «узнать», что вычисление $C_k(k)$ не завершается, даже если процедура отыскания $C_k(k)$ с помощью $A$ является целиком и полностью алгоритмической. Мы, разумеется, могли бы *сообщить* роботу о том, что вычисление $C_k(k)$ и в самом деле не завершается (воспользовавшись для установления этого факта собственными пониманием и интуицией), однако, если робот примет это утверждение на «веру», ему придется изменить свои собственные правила, присоединив полученную новую истину к тем, что он уже «знает». Мы можем пойти еще дальше и каким-либо способом сообщить нашему роботу о том, что для получения новых истин на основании старых ему, помимо прочего, необходимо «знать» и общую вычислительную процедуру отыскания $C_k(k)$ посредством алгоритма $A$. К запасу «знаний» робота можно добавить все, что является вполне

определенным и вычислительным по своей природе. Однако в результате у нас появляется *новый* алгоритм «$A$», и доказательство Гёделя следует применять уже к нему, а не к старому $A$. Иначе говоря, везде вместо старого $A$ нам следовало бы использовать новый «$A$», поскольку менять алгоритм посреди доказательства есть не что иное, как жульничество. Таким образом, как мы видим, изъян возражения **Q6** очень похож на рассмотренный выше изъян **Q5**. В нашем *reductio ad absurdum* мы полагаем, что алгоритм $A$ (под которым понимается некая познаваемая и обоснованная процедура для установления факта незавершаемости вычислений) в действительности представляет собой *всю совокупность* известных математикам подобных процедур, из чего и следует противоречие. Попытку введения еще одной вычислительной процедуры для установления истины — процедуры, не содержащейся в $A$, — *после* того как мы договорились, что $A$ представляет собой всю их совокупность, я расцениваю как жульничество.

Беда нашего злосчастного робота в том, что, не обладая каким бы то ни было *пониманием* гёделевской процедуры, он не располагает ни одним надежным и независимым способом установления истины — истину ему сообщаем мы. (Эта проблема, вообще говоря, не имеет никакого отношения к вычислительным аспектам доказательства Гёделя.) Для того чтобы достичь чего-то большего, ему, как и всем нам, необходимо понимание смысла операций, которые ему велено выполнять. Если такого понимания нет, то он вполне может «знать» (ошибочно), что вычисление $C_k(k)$ *завершается*, а вовсе не наоборот. Заключение (ошибочное) «вычисление $C_k(k)$ завершается» выводится точно так же алгоритмически, как и заключение (правильное) «вычисление $C_k(k)$ не завершается». Таким образом, дело вовсе не в алгоритмическом характере этих операций, а в том, что для различения между алгоритмами, приводящими к истинным заключениям, и теми, что приводят к заключениям ложным, наш робот нуждается в способности выносить достоверные *суждения об истинности*. Далее, на данной стадии рассуждения, мы все еще допускаем возможность того, что процесс «понимания» представляет собой некую разновидность алгоритмической деятельности, которая не содержится ни в одной из точно заданных и «заведомо» обоснованных процедур типа $A$. Например, понимание может осуществляться посредством выполнения какого-то



необоснованного или непознаваемого алгоритма. В дальнейшем (см. главу 3) я попробую убедить читателя в том, что в действительности понимание вообще не является алгоритмической деятельностью. На настоящий же момент нас интересуют всего лишь строгие следствия из доказательства Гёделя–Тьюринга, а на них возможность получения вычисления $C_k(k)$ из процедуры $A$ вычислительным путем никоим образом не влияет.

**Q7. Общая совокупность результатов, полученных всеми когда-либо жившими математиками, плюс совокупность результатов, которые будут получены всеми математиками за последующую, скажем, тысячу лет, — имеет конечную величину и может уместиться в банках памяти соответствующего компьютера. Такой компьютер, естественно, *способен* без особого труда воспроизвести все эти результаты, и, тем самым, повести себя (внешне) как математик-человек — что бы ни утверждало по этому поводу гёделевское доказательство.**

Несмотря на кажущуюся логичность этого утверждения, здесь упущен из виду один очень существенный момент, а именно: способ, посредством которого мы (или компьютеры) определяем, какие математические утверждения истинны, а какие — ложны. (Во всяком случае, на простое *хранение* математических утверждений способны и системы, гораздо менее сложные, нежели универсальный компьютер, — например, фотоаппараты.) Принцип использования компьютера в **Q7** совершенно не учитывает критического вопроса о наличии у этого самого компьютера способности *суждения об истинности*. С равным успехом можно вообразить и компьютеры, в памяти которых не содержится ничего, кроме перечня абсолютно ложных математических «теорем», либо случайным образом перемешанных истинных и ложных утверждений. Откуда мы узнаем, какому компьютеру можно доверять? Я отнюдь не утверждаю, что эффективное моделирование результатов сознательной интеллектуальной деятельности человека (в данном случае, в области математики) абсолютно невозможно, поскольку по одной лишь чистой случайности компьютер может «умудриться» сделать все правильно, пусть и не обладая каким бы то ни было пониманием. Однако шансы на это до абсурдного малы, в то время как те вопросы, на которые мы

здесь пытаемся найти ответ (например, каким таким образом мы *определяем*, что вот это математическое утверждение истинно, а вот это — ложно?), в возражении **Q7** и вовсе не затрагиваются.

С другой стороны, **Q7** все же напоминает об одном более существенном соображении. Имеет ли непосредственное отношение к нашему исследованию обсуждение бесконечных структур (*всех* натуральных чисел или *всех* вычислений), если учесть, что совокупность всех результатов, полученных на тот или иной момент времени всеми людьми и компьютерами, имеет *конечную* величину? В следующем комментарии мы рассмотрим этот безусловно важный вопрос отдельно.

**Q8. Незавершающиеся вычисления суть идеализированные математические конструкции, по определению бесконечные. Вряд ли подобные вопросы могут иметь сколько-нибудь непосредственное отношение к изучению конечных физических объектов — таких, как компьютеры или мозг.**

Все верно: рассуждая в идеализированном ключе о машинах Тьюринга, незавершающихся вычислениях и т. п., мы рассматривали бесконечные (потенциально) процессы, тогда как в случае людей или компьютеров нам приходится иметь дело с системами *конечными*. И, разумеется, применяя подобные идеализированные доказательства к реальным и конечным физическим объектам, следует быть готовыми к тому, что такая операция непременно окажется связанной с теми или иными ограничениями и оговорками. Однако, как выясняется, учет конечной природы реальных объектов не изменяет сколько-нибудь существенно сути доказательства Гёделя—Тьюринга. Нет ничего странного в том, что мы *рассуждаем* об идеализированных вычислениях, обосновываем те или иные умозаключения и выводим, математически, их теоретические ограничения. Можно, к примеру, обсуждать в абсолютно конечных терминах вопрос о том, существует ли нечетное число, являющееся суммой двух четных чисел, или существует ли натуральное число, не являющееся суммой четырех квадратов (как в приведенных выше задачах **(C)** и **(B)**), нисколько не смущаясь тем, что при рассмотрении этих вопросов мы неявно учитываем бесконечное множество *всех* натуральных чисел. Мы имеем полное право рассуждать о незавершающихся вычислениях (или машинах Тьюринга вообще) как о *математических*



структурах, пусть и не в силах создать на практике бесконечно работающую машину Тьюринга. (Отметим, в частности, что действие машины Тьюринга, занятой поисками нечетного числа, являющегося суммой двух четных чисел, строго говоря, практически реализовать невозможно, так как ее детали износятся гораздо раньше, чем минет вечность.) Описание любого единичного вычисления (или действия машины Тьюринга) — задача вполне конечная, а вопрос о том, завершится ли в конечном итоге это вычисление, можно полагать вполне определенным. *Сначала* мы доводим до логического завершения теоретические рассуждения, связанные с теми или иными идеализированными вычислениями, и лишь *затем* пытаемся разглядеть, каким образом наши рассуждения применимы к конечным физическим системам — таким, как реально существующие компьютеры или люди.

Ограничения конечного характера могут быть обусловлены либо тем, что (i) описание конкретного рассматриваемого вычисления оказывается слишком громоздким (т. е. число $n$ в $C_n$ или пара чисел $q$, $n$ в $C_q(n)$ оказываются слишком велики для того, чтобы их мог описать человек или реально существующий компьютер), либо тем, что (ii) при внешней простоте описания вычисление, тем не менее, требует для своего выполнения чрезмерно много времени, в результате чего может показаться, что оно не завершается вовсе, хотя теоретически данное вычисление должно в конечном счете завершиться. На деле же, как мы вскоре убедимся, выясняется, что из этих двух условий сколько-нибудь существенное влияние на наши рассуждения оказывает только (i), да и оно не так уж и велико. Незначительность фактора (ii), быть может, покажется вам удивительной. Существует множество относительно простых вычислений, которые в конечном счете завершаются, однако точки их завершения путем прямого вычисления не способен достичь ни один потенциально возможный компьютер. Рассмотрим, например, следующую задачу: «распечатать последовательность из $2^{2^{65536}}$ единиц, после чего остановиться». (В § 3.26 будут предложены еще несколько подобных примеров, гораздо более интересных с математической точки зрения.) Вопрос о завершаемости того или иного вычисления не следует решать путем прямого вычисления: этот метод зачастую оказывается крайне неэффективным.

Для того чтобы выяснить, каким образом ограничения (i) или (ii) могут повлиять на наши гёделевские рассуждения, пройдемся

еще раз по соответствующим частям доказательства. В соответствии с ограничением (i), вместо бесконечного ряда вычислений, мы располагаем рядом *конечным*:

$$C_0, C_1, C_2, C_3, \ldots, C_Q,$$

где предполагается, что число $Q$ задает наиболее громоздкое вычисление, какое способен выполнить наш компьютер или человек. В случае с человеком вышеприведенное утверждение можно счесть несколько туманным. Впрочем, в настоящий момент нас не особенно заботит точное определение числа $Q$. (Вопрос о туманности утверждений, касающихся человеческих способностей, будет рассмотрен ниже, в комментарии к возражению **Q13** в §2.10.) Кроме того, можно предположить, что, попытавшись применить упомянутые вычисления к какому-то конкретному натуральному числу $n$, мы обнаружим, что значение $n$ ограничено некоторой фиксированной величиной $N$, поскольку наш компьютер (или человек) оказывается не способен работать с числами, превышающими $N$. (Строго говоря, следует учесть и возможность того, что число $N$ не является фиксированным, но зависит от того или иного конкретного вычисления $C_q$, т. е. $N$ может зависеть от $q$. Однако этот факт не влияет на наши рассуждения сколько-нибудь существенным образом.)

Как и ранее, мы рассматриваем некий обоснованный алгоритм $A(q, n)$, завершение выполнения которого равносильно доказательству того, что вычисление $C_q(n)$ не завершается. Несмотря на то, что, в соответствии с ограничением (i), рассмотрению подлежат только значения $q$, не превышающие $Q$, и только значения $n$, не превышающие $N$, мы, говоря об «обоснованности», в действительности имеем в виду, что алгоритм $A$ должен быть обоснованным для *всех* значений $q$ и $n$, независимо от их величины. (Таким образом, можно видеть, что правила, реализуемые в алгоритме $A$, являются точными *математическими* правилами, в отличие от правил приближенных, работающих только в силу того или иного практического ограничения, налагаемого на «реально осуществимые» вычисления.) Более того, утверждая, что «вычисление $C_q(n)$ не завершается», мы имеем в виду, что это вычисление *действительно* не завершается, а не то, что это вычисление просто-напросто оказывается слишком громоздким для того, чтобы его мог выполнить наш компьютер или человек, как предусматривает ограничение (ii).



Вспомним, что утверждение **(H)** гласит:

> Если завершается вычисление $A(q, n)$, то вычисление $C_q(n)$ не завершается.

Принимая во внимание ограничение (ii), можно было бы предположить, что алгоритм $A$ оказывается не слишком эффективен при установлении факта незавершаемости очередного вычисления, поскольку сам он состоит из большего количества шагов, чем способен выполнить компьютер или человек. Однако, как выясняется, для нашего доказательства этот факт не имеет никакого значения. Мы намерены отыскать некое вычисление $A(k, k)$, которое не завершается вообще. Для нас абсолютно неважно, что в некоторых других случаях, когда вычисление $A$ *действительно* завершается, мы не можем об этом узнать, так как не в состоянии дождаться этого самого завершения.

Далее, как и в равенстве **(J)**, мы вводим натуральное число $k$, при котором вычисление $A(n, n)$ совпадает с вычислением $C_k(n)$ для всех $n$:

$$A(n, n) = C_k(n).$$

Следует, впрочем, рассмотреть еще предусматриваемую ограничением (i) возможность того, что упомянутое число $k$ окажется больше $Q$. В случае какого-нибудь невообразимо сложного вычисления $A$ такая ситуация вполне возможна, однако только при условии, что это $A$ уже начинает приближаться к верхней границе допустимой сложности (в смысле количества двоичных знаков в его описании в формате машины Тьюринга), с которой может работать наш компьютер или человек. Это обусловлено тем, что вычисление, получающее значение $k$ из описания вычисления $A$ (например, в формате машины Тьюринга), — вещь достаточно простая и может быть задана в явном виде (как уже было показано в комментарии к **Q6**).

Вообще говоря, для того чтобы поставить в тупик алгоритм $A$, нам необходимо лишь вычисление $C_k(k)$ — подставляя в **(H)** равенство $n = k$, получаем утверждение **(L)**:

> Если завершается вычисление $A(k, k)$, то вычисление $C_k(k)$ не завершается.

Поскольку $A(k, k)$ совпадает с $C_k(k)$, наше доказательство показывает, что, хотя данное конкретное вычисление $C_k(k)$ никогда

не завершается, посредством алгоритма $A$ мы этот факт установить не в состоянии, даже если бы упомянутый алгоритм мог выполняться гораздо дольше любого предела, налагаемого на него в соответствии с ограничением (ii). Вычисление $C_k\,(k)$ задается только введенным ранее числом $k$, и, при условии, что $k$ не превышает ни $Q$, ни $N$, это вычисление и в самом деле в состоянии выполнить наш компьютер или человек — то есть в состоянии *начать*. Довести его до завершения невозможно в любом случае, поскольку это вычисление просто-напросто не завершается!

А может ли число $k$ оказаться больше $Q$ или $N$? Такое возможно лишь в том случае, когда для описания $A$ требуется так много знаков, что даже совсем небольшое увеличение их количества выводит задачу за пределы возможностей нашего компьютера или человека. При этом, поскольку мы знаем об обоснованности алгоритма $A$, мы *знаем* и о том, что рассматриваемое вычисление $C_k\,(k)$ не завершается, даже если реальное выполнение этого вычисления представляет для нас проблему. Соображение (i), однако, предполагает и возможность того, что вычисление $A$ окажется столь колоссально сложным, что одно лишь его описание вплотную приблизится к доступному воображению человека пределу сложности, а сравнительно малое увеличение количества составляющих его знаков даст в результате вычисление, превосходящее всякое человеческое понимание. Что бы мы о подобной возможности ни думали, я все же считаю, что любой столь впечатляющий набор реализуемых в нашем гипотетическом алгоритме $A$ вычислительных правил окажется, вне всякого сомнения, настолько сложным, что мы не в состоянии будем *знать* наверняка, является ли он *обоснованным*, даже если нам будут точно известны все эти правила по отдельности. Таким образом, наше прежнее заключение остается в силе: при установлении математических истин мы *не* применяем *познаваемо обоснованные* наборы алгоритмических правил.

Не помешает несколько более подробно остановиться на сравнительно незначительном увеличении сложности, сопровождающем переход от $A$ к $C_k\,(k)$. Помимо прочего, это существенно поможет нам в нашем дальнейшем исследовании (в §§ 3.19 и 3.20). В Приложении А (с. 193) предложено явное описание вычисления $C_k\,(k)$ в виде предписаний для машины Тьюринга, рассмотренных в НРК (глава 2). Согласно этим предписаниям, под обозначением $T_m$ понимается «$m$-я машина Тьюринга». Для



большего удобства и упрощения рассуждений здесь мы также будем пользоваться этим обозначением вместо «$C_m$», в частности, для определения *степени сложности* вычислительной процедуры или отдельного вычисления. В соответствии с вышесказанным, определим степень сложности $\mu$ машины Тьюринга $T_m$ как количество знаков в двоичном представлении числа $m$ (см. НРК, с. 39); при этом степень сложности некоторого вычисления $T_m\,(n)$ определяется как большее из двух чисел $\mu$ и $\nu$, где $\nu$ — количество двоичных знаков в представлении числа $n$. Рассмотрим далее приведенное в Приложении А явное предписание для составления вычисления $C_k\,(k)$ на основании алгоритма $A$, заданного в упомянутых спецификациях машины Тьюринга. Полагая степень сложности $A$ равной $\alpha$, находим, что степень сложности явного вычисления $C_k\,(k)$ не превышает числа $\alpha + 210\,\log_2\,(\alpha + 336)$ — а это число, в свою очередь, оказывается лишь очень ненамного больше собственно $\alpha$, да и то только тогда, когда число $\alpha$ очень велико.

В вышеприведенных общих рассуждениях имеется один потенциально спорный момент. В самом деле, какой смысл рассматривать вычисления, слишком сложные даже для того, чтобы просто их записать, или те, что, будучи записанными, возможно, потребуют на свое действительное выполнение промежуток времени, гораздо больший предполагаемого возраста нашей Вселенной, даже при условии, что каждый шаг такого вычисления будет производиться за самую малую долю секунды, какая еще допускает протекание каких бы то ни было физических процессов? Упомянутое выше вычисление — то, результатом которого является последовательность из $2^{2^{65536}}$ единиц и которое завершается лишь *после* выполнения этой задачи, — представляет собой как раз такой пример; при этом позицию математика, позволяющего себе утверждать, что данное вычисление является незавершающимся, можно охарактеризовать как крайне нетрадиционную. Однако в математике существуют и некоторые другие точки зрения, пусть и не до *такой* степени нетрадиционные, — но все же решительно презирающие всяческие условности, — согласно которым известная доля здорового скептицизма в отношении вопроса об абсолютной математической истинности идеализированных математических утверждений отнюдь не помешает. На некоторые из них, безусловно, стоит хотя бы мельком взглянуть.

**Q9. Точка зрения, известная как *интуиционизм*, не позволяет сделать вывод о непременной завершаемости вычисления на определенном этапе на том лишь основании, что бесконечное продолжение этого вычисления приводит к противоречию; бытуют в математике и иные точки зрения сходного характера — например, «конструктивизм» и «финитизм». Не окажется ли гёделевское доказательство спорным, будучи рассмотрено с этих позиций?**

В своем гёделевском доказательстве (в частности, в утверждении **(M)**) я использовал аргумент следующего вида: «Допущение о ложности $X$ приводит к противоречию; следовательно, утверждение $X$ истинно». Под «$X$» в данном случае следует понимать утверждение: «Вычисление $C_k(k)$ не завершается». Это рассуждение относится к типу *reductio ad absurdum*; что же касается доказательства Гёделя в целом, то оно и в самом деле построено именно таким образом. Направление же в математике, называемое «интуиционизмом» (у истоков которого стоял голландский математик Л. Э. Я. Брауэр; см. [223] и НРК, с. 113–116), отрицает возможность построения обоснованного доказательства на основе *reductio ad absurdum*. Интуиционизм возник приблизительно в 1912 году как реакция на некоторые сформировавшиеся к концу девятнадцатого — началу двадцатого века математические тенденции, суть которых сводится к следующему: математический объект можно полагать «существующим» даже в тех случаях, когда нет никакой возможности этот объект так или иначе воплотить в действительности. А надо сказать, что слишком вольное применение крайне расплывчатой концепции математического существования и впрямь приводит порой к весьма неприятным противоречиям. Самый известный пример такого противоречия связан с парадоксальным «множеством всех множеств, не являющихся членами самих себя» Бертрана Рассела. (Если множество Рассела является членом самого себя, то оно таковым не является; если же оно членом самого себя не является, то оно им, как ни странно, является! Подробнее см. §3.4 и НРК, с. 101.) Дабы противостоять общей тенденции, в рамках которой могут считаться «существующими» весьма вольно определенные математические объекты, интуиционисты полагают необоснованным математическое рассуждение, позволяющее



делать вывод о существовании того или иного математического объекта на основании одной лишь противоречивости его несуществования. Доказательство существования объекта посредством *reductio ad absurdum* не дает абсолютно никаких оснований полагать, что упомянутый объект действительно можно построить.

Каким же образом запрет на применение *reductio ad absurdum* может повлиять на наше гёделевское доказательство? Вообще говоря, совсем не может, по той простой причине, что *reductio ad absurdum* мы применяем, если можно так выразиться, наоборот, то есть противоречие в нашем случае выводится из допущения, что нечто *существует*, а не из обратного допущения. С интуиционистской точки зрения все выглядит совершенно законно: мы заключаем, что объект *не* существует, на том основании, что противоречие возникает как раз из допущения о существовании этого самого объекта. Предложенное мною гёделевское доказательство, по сути своей, является в интуиционистском смысле абсолютно приемлемым. (См. [223], с. 492.)

Аналогичные рассуждения применимы и ко всем прочим «конструктивистским» или «финитистским» направлениям в математике, о каких мне известно. Комментарий к возражению **Q8** демонстрирует, что даже та точка зрения, согласно которой последовательность натуральных чисел нельзя считать «на самом деле» бесконечной, не освобождает нас от неизбежного вывода: для установления математической истины мы таки не пользуемся познаваемо обоснованными алгоритмами.

## 2.7. Некоторые более глубокие математические соображения

Для того чтобы лучше разобраться в значении гёделевского доказательства, полезно будет вспомнить, с какой, собственно, целью оно было первоначально предпринято. На рубеже веков ученые, деятельность которых была связана с фундаментальными математическими принципами, столкнулись с весьма серьезными проблемами. В конце XIX века — в значительной степени благодаря глубоко оригинальным математическим трудам Георга Кантора (с «диагональным доказательством» которого мы уже познакомились) — математики получили в распоряжение эффективные методы доказательства некоторых наиболее фунда-

ментальных своих результатов, основанные на свойствах *бесконечных множеств*. Однако с этими преимуществами оказались связаны и не менее фундаментальные трудности, проистекающие из чересчур вольного обращения с концепцией бесконечного множества. Особо отметим парадокс Рассела (на который я уже ссылался в комментарии к **Q9**, см. также § 3.4 — Кантор о нем также упоминает), обозначивший некоторые препятствия, подстерегающие склонных к опрометчивым умозаключениям. Тем не менее, все понимали, что если вопрос о допустимости тех или иных методов рассуждения продумать с достаточной тщательностью, то можно добиться очень и очень впечатляющих математических результатов. Проблема, по всей видимости, сводилась к отысканию способа, посредством которого можно было бы в каждом конкретном случае абсолютно *точно* определить, была ли соблюдена при выборе метода рассуждения «достаточная тщательность».

Одной из главных фигур движения, поставившего перед собой цель достичь этой точности, был великий математик Давид Гильберт. Движение окрестили *формализмом*; в соответствии с его основополагающим принципом, следовало однозначно определить все допустимые методы математического рассуждения в пределах той или иной конкретной области раз и навсегда, включая и те, что связаны с понятием бесконечного множества. Такая совокупность правил и математических утверждений называется *формальной системой*. После того как определены правила формальной системы $\mathbb{F}$, решение вопроса о корректности применения этих правил — количество которых непременно является конечным[6] — сводится к элементарной механической проверке. Разумеется, если мы хотим, чтобы любой выводимый с помощью таких правил результат мог считаться действительно *истинным*, нам придется присвоить им всем статус вполне допустимых и об-

[6]Представление некоторых формальных систем включает в себя *бесконечное* количество аксиом (они описываются через посредство структур, называемых «схемами аксиом»), однако, чтобы оставаться «формальной» в том смысле, какой вкладываю в это понятие я, система должна быть выразима в каком-то конечном виде — например, упомянутая система с бесконечным количеством аксиом должна порождаться конечным набором вычислительных правил. Это вполне возможно, и именно так и обстоит дело со стандартными формальными системами, которые применяются в математических доказательствах, — одной из таких систем является, например, знаменитая «формальная система Цермело–Френкеля» **ZF**, описывающая традиционную теорию множеств.



основанных форм математического рассуждения. Однако некоторые из рассматриваемых правил могут подразумевать какие-либо манипуляции с бесконечными множествами, и в этом случае математическая интуиция, подсказывающая нам, какие методы рассуждения допустимы, а какие нет, может оказаться и не достойной абсолютного доверия. Сомнения в этой связи как нельзя более уместны, учитывая несоответствия, возникающие при столь вольном обращении с бесконечными множествами, что допустимым становится даже парадоксальное «множество всех множеств, не являющихся членами самих себя» Бертрана Рассела. Правила системы $\mathbb{F}$ не должны допускать существования «множества» Рассела, но где же, в таком случае, следует провести границу? Вообще запретить применение бесконечных множеств было бы слишком строгим ограничением (обычное евклидово пространство, например, содержит бесконечное множество точек, да и множество натуральных чисел является бесконечным); кроме того, существуют же формальные системы, абсолютно в этом смысле удовлетворительные (поскольку в их рамках не допускается, к примеру, формулировать сущности, подобные «множеству» Рассела), применяя которые можно получить бóльшую часть необходимых математических результатов. Откуда нам знать, каким из этих формальных систем можно верить, а каким нельзя?

Рассмотрим подробнее одну такую формальную систему $\mathbb{F}$; для математических утверждений, которые можно получить с помощью правил системы $\mathbb{F}$, введем обозначение ИСТИННЫЕ, а для утверждений, *отрицания* которых выводятся из того же источника (т. е. утверждения, обратные рассматриваемым), — обозначение ЛОЖНЫЕ. Любое утверждение, которое можно сформулировать в рамках системы $\mathbb{F}$, но которое не является в этом смысле ни ИСТИННЫМ, ни ЛОЖНЫМ, будем полагать НЕРАЗРЕШИМЫМ. Кто-то, возможно, сочтет, что поскольку на деле может оказаться «бессмысленным» и само понятие бесконечного множества, то, по всей видимости, нельзя абсолютно осмысленно говорить ни об истинности, ни о ложности относящихся к ним утверждений. (Это мнение применимо по крайней мере к некоторым разновидностям бесконечных множеств, если не ко всем.) Если придерживаться такой точки зрения, то нет особой разницы, какие именно утверждения о бесконечных множествах (некоторых разновидностей) оказываются ИСТИННЫМИ, а какие —

ЛОЖНЫМИ, лишь бы не вышло так, что одно утверждение получится ИСТИННЫМ *и* ЛОЖНЫМ одновременно, т. е. система $\mathbb{F}$ должна все же быть *непротиворечивой*. Собственно говоря, в этом и состоит суть истинного *формализма*, а в отношении формальной системы $\mathbb{F}$ первостепенно важно знать лишь следующее: (a) является ли она *непротиворечивой* и (b) является ли она *полной*. Система $\mathbb{F}$ называется *полной*, если любое математическое утверждение, должным образом сформулированное в рамках $\mathbb{F}$, всегда оказывается либо ИСТИННЫМ, либо ЛОЖНЫМ (т. е. НЕРАЗРЕШИМЫХ утверждений система $\mathbb{F}$ не содержит).

Для строгого формалиста вопрос о том, является ли то или иное утверждение о бесконечных множествах *действительно истинным* в сколько угодно абсолютном смысле, не обязательно имеет смысл и, уж конечно же, не имеет никакого существенного отношения к процедурам формалистской математики. Таким образом, поиски абсолютной математической истины в отношении утверждений, связанных с упомянутыми бесконечными величинами, заменяются стремлением продемонстрировать непротиворечивость и полноту соответствующих формальных систем. Какие же математические правила допустимо использовать для такой демонстрации? Достойные доверия, прежде всего, причем формулировка этих правил ни в коем случае не должна основываться на сомнительных рассуждениях с привлечением слишком вольно определяемых бесконечных множеств (типа множества Рассела). Была надежда на то, что в рамках некоторых сравнительно простых и очевидно обоснованных формальных систем (например, такой достаточно элементарной системы, как *арифметика Пеано*) отыщутся логические процедуры, которых будет достаточно для того, чтобы доказать непротиворечивость других, более сложных, формальных систем — скажем, системы $\mathbb{F}$, — непротиворечивость которых уже не столь бесспорна и в рамках которых допускаются формальные рассуждения об очень «больших» бесконечных множествах. Если принять философию формалистов, то подобное доказательство непротиворечивости для $\mathbb{F}$, как минимум, даст основание для использования методов рассуждения, допустимых в рамках системы $\mathbb{F}$. Затем можно доказывать математические теоремы, применяя концепцию бесконечных множеств тем или иным непротиворечивым образом, а может, удастся и вовсе избавиться от необходимости отвечать на вопрос о реальном «смысле» таких множеств. Более того, если



удастся показать, что система $\mathbb{F}$ является еще и полной, то можно будет вполне резонно счесть, что эта система действительно содержит абсолютно *все* допустимые математические процедуры, т. е. представляет собой, в некотором смысле, *полное* описание математического аппарата рассматриваемой области.

Однако в 1930 году (публикация состоялась в 1931) Гёдель взорвал свою «бомбу», раз и навсегда показав, что идеал формалистов принципиально недостижим. Он продемонстрировал, что не может существовать формальной системы $\mathbb{F}$, которая была бы одновременно и непротиворечивой (в некоем «сильном» смысле, который мы рассмотрим в следующем разделе), и полной, — при условии, что $\mathbb{F}$ считается достаточно мощной, чтобы сочетать в себе формулировки утверждений обычной арифметики и стандартную логику. Таким образом, теорема Гёделя справедлива для таких систем $\mathbb{F}$, в рамках которых арифметические утверждения типа теоремы Лагранжа и гипотезы Гольдбаха (см. §2.3) формулируются как утверждения математические.

В дальнейшем мы будем рассматривать только те формальные системы, которые являются достаточно обширными, чтобы содержать в себе необходимые для действительной формулировки теоремы Гёделя арифметические операции (а также, в случае нужды, и операции какой угодно машины Тьюринга; см. ниже). Говоря о какой-либо формальной системе $\mathbb{F}$, я обычно буду *подразумевать*, что она действительно достаточно обширна в этом смысле. Это допущение не отразится на наших рассуждениях сколько-нибудь существенным образом. (Тем не менее, рассматривая формальные системы в таком контексте, я, для пущей ясности, буду иногда снабжать их эпитетом «достаточно обширная» или иным подобным.)

## 2.8. Условие ω-непротиворечивости

Наиболее известная форма теоремы Гёделя гласит, что формальная система $\mathbb{F}$ (достаточно обширная) не может быть одновременно полной и непротиворечивой. Это не совсем та знаменитая «теорема о неполноте», которую Гёдель первоначально представил на конференции в Кенигсберге (см. §§2.1 и 2.7), а ее несколько более сильный вариант, который был позднее получен американским логиком Дж. Баркли Россером (1936). По своей сути, первоначальный вариант теоремы Гёделя оказывается эквивалентен утверждению, что система $\mathbb{F}$ не может быть

одновременно полной и *ω-непротиворечивой*. Условие же ω-непротиворечивости несколько строже, нежели условие непротиворечивости обыкновенной. Для объяснения его смысла нам потребуется ввести некоторые новые обозначения. В систему обозначений формальной системы $\mathbb{F}$ необходимо включить символы некоторых логических операций. Нам, в частности, потребуется символ, выражающий *отрицание* («не»); можно выбрать для этого символ «∼». Таким образом, если $Q$ есть некое высказывание, формулируемое в рамках $\mathbb{F}$, то последовательность символов $\sim Q$ означает «не $Q$». Нужен также символ, означающий «для всех [натуральных чисел]» и называемый *квантор общности*; он имеет вид «∀». Если $P(n)$ есть некое высказывание, зависящее от натурального числа $n$ (т. е. $P$ представляет собой так называемую *пропозициональную функцию*), то строка символов $\forall n\,[P(n)]$ означает «для всех натуральных чисел $n$ высказывание $P(n)$ справедливо». Например, если высказывание $P(n)$ имеет вид «число $n$ можно выразить в виде суммы квадратов трех чисел», то запись $\forall n\,[P(n)]$ означает «любое натуральное число является суммой квадратов трех чисел», — что, вообще говоря, ложно (хотя, если мы заменим «трех» на «четырех», то это же утверждение станет истинным). Такие символы можно записывать в самых различных сочетаниях; в частности, строка символов

$$\sim \forall n\,[P(n)]$$

выражает *отрицание* того, что высказывание $P(n)$ справедливо для всех натуральных чисел $n$.

Условие же ω-непротиворечивости гласит, что если высказывание $\sim \forall n\,[P(n)]$ можно доказать с помощью методов формальной системы $\mathbb{F}$, то это еще *не* означает, что в рамках этой самой системы непременно доказуемы *все* утверждения

$$P(0), P(1), P(2), P(3), P(4), \ldots.$$

Отсюда следует, что если формальная система $\mathbb{F}$ не является ω-непротиворечивой, мы оказываемся в аномальной ситуации, когда для некоторого $P$ оказывается доказуемой истинность всех высказываний $P(0)$, $P(1)$, $P(2)$, $P(3)$, $P(4)$, ...; и *одновременно с этим* можно доказать и то, что *не* все эти высказывания истинны! Безусловно, ни одна заслуживающая доверия формальная система подобного безобразия допустить не может. Поэтому



если система $\mathbb{F}$ является *обоснованной*, то она непременно будет и ω-непротиворечивой.

В дальнейшем утверждения «формальная система $\mathbb{F}$ является непротиворечивой» и «формальная система $\mathbb{F}$ является ω-непротиворечивой» я буду обозначать, соответственно, символами «$G(\mathbb{F})$» и «$\Omega(\mathbb{F})$». В сущности (если полагать систему $\mathbb{F}$ достаточно обширной), сами утверждения $G(\mathbb{F})$ и $\Omega(\mathbb{F})$ формулируются как операции этой системы. Согласно знаменитой теореме Гёделя о неполноте, утверждение $G(\mathbb{F})$ *не является теоремой* системы $\mathbb{F}$ (т. е. его нельзя доказать с помощью процедур, допустимых в рамках системы $\mathbb{F}$); не является теоремой и утверждение $\Omega(\mathbb{F})$ — если, разумеется, система $\mathbb{F}$ *действительно* непротиворечива. Несколько более строгий вариант теоремы Гёделя, сформулированный позднее Россером, гласит, что если система $\mathbb{F}$ непротиворечива, то утверждение $\sim G(\mathbb{F})$ также не является теоремой этой системы. В оставшейся части этой главы я буду формулировать свои доводы не столько исходя из утверждения $\Omega(\mathbb{F})$, сколько на основе более привычного нам $G(\mathbb{F})$, хотя для большей части наших рассуждений в равной степени сгодится любое из них. (В некоторых наиболее явных аргументах главы 3 я буду иногда обозначать через «$G(\mathbb{F})$» конкретное утверждение «вычисление $C_k(k)$ не завершается» (см. §2.5); надеюсь, никто не сочтет это слишком большой вольностью с моей стороны.)

В большей части предлагаемых рассуждений я не стану проводить четкую границу между непротиворечивостью и ω-непротиворечивостью, однако тот вариант теоремы Гёделя, что представлен в §2.5, по сути, гласит, что если формальная система $\mathbb{F}$ непротиворечива, то она не может быть полной, так как не может включать в себя в качестве теоремы утверждение $G(\mathbb{F})$. Здесь я всего этого демонстрировать не буду (интересующиеся же могут обратиться к [223]). Вообще говоря, для того чтобы эту форму гёделевского доказательства можно было свести к доказательству в моей формулировке, система $\mathbb{F}$ должна содержать в себе нечто большее, нежели просто «арифметику и обыкновенную логику». Необходимо, чтобы система $\mathbb{F}$ была обширной настолько, чтобы включать в себя действия любой *машины Тьюринга*. Иначе говоря, среди утверждений, корректно формулируемых с помощью символов системы $\mathbb{F}$, должны присутствовать утверждения типа: «Такая-то машина Тьюринга, оперируя над натуральным числом $n$, дает на выходе натуральное число $p$».

Более того, имеется теорема (см. [223], главы 11 и 13), согласно которой так оно само собой и получается, если, помимо обычных арифметических операций, система $\mathbb{F}$ содержит следующую операцию (так называемую $\mu$-операцию, или операцию минимизации): «найти наименьшее натуральное число, обладающее таким-то арифметическим свойством». Вспомним, что в нашем первом вычислительном примере, **(A)**, предложенная процедура действительно позволяла отыскать *наименьшее* число, не являющееся суммой трех квадратов. То есть, вообще говоря, право на подобные вещи за вычислительными процедурами следует сохранить. С другой стороны, именно благодаря *этой* их особенности мы и сталкиваемся с вычислениями, которые принципиально не завершаются, — например, вычисление **(B)**, где мы пытаемся отыскать наименьшее число, не являющееся суммой *четырех* квадратов, а такого числа в природе не существует.

## 2.9. Формальные системы и алгоритмическое доказательство

В предложенной мною формулировке доказательства Гёделя—Тьюринга (см. § 2.5) говорится только о «вычислениях» и ни словом не упоминается о «формальных системах». Тем не менее, между этими двумя концепциями существует очень тесная связь. Одним из существенных свойств формальной системы является непременная необходимость существования алгоритмической (т. е. «вычислительной») процедуры $F$, предназначенной для *проверки* правильности применения правил этой системы. Если, в соответствии с правилами системы $\mathbb{F}$, некое высказывание является ИСТИННЫМ, то вычисление $F$ этот факт установит. (Для достижения этого результата вычисление $F$, возможно, «просмотрит» все возможные последовательности строк символов, принадлежащих «алфавиту» системы $\mathbb{F}$, и успешно завершится, обнаружив заключительной строкой искомое высказывание $P$; при этом любые сочетания строк символов являются, согласно правилам системы $\mathbb{F}$, допустимыми.)

Напротив, располагая некоторой *заданной* вычислительной процедурой $E$, предназначенной для установления истинности определенных математических утверждений, мы можем построить формальную систему $\mathbb{E}$, которая эффективно выражает



как ИСТИННЫЕ все те истины, что можно получить с помощью процедуры $E$. Имеется, впрочем, и небольшая оговорка: как правило, формальная система должна содержать стандартные логические операции, однако заданная процедура $E$ может оказаться недостаточно обширной, чтобы непосредственно включить и их. Если сама заданная процедура $E$ не содержит этих элементарных логических операций, то при построении системы $\mathbb{E}$ уместно будет присоединить их к $E$ с тем, чтобы ИСТИННЫМИ положениями системы $\mathbb{E}$ оказались не только утверждения, получаемые непосредственно из процедуры $E$, но и утверждения, являющиеся элементарными логическими следствиями утверждений, получаемых непосредственно из $E$. При таком построении система $\mathbb{E}$ не будет строго эквивалентна процедуре $E$, но вместо этого приобретет несколько большую мощность.

(Среди таких логических операций могут, к примеру, оказаться следующие: «если $P\&Q$, то $P$»; «если $P$ и $P \Rightarrow Q$, то $Q$»; «если $\forall x\,[P(x)]$, то $P(n)$»; «если $\sim \forall x\,[P(x)]$, то $\exists x\,[\sim P(x)]$» и т. п. Символы «&», «$\Rightarrow$», «$\forall$», «$\exists$», «$\sim$» означают здесь, соответственно, «и», «следует», «для всех [натуральных чисел]», «существует [натуральное число]», «не»; в этот ряд можно включить и некоторые другие аналогичные символы.)

Поставив перед собой задачу построить на основе процедуры $E$ формальную систему $\mathbb{E}$, мы можем начать с некоторой в высшей степени фундаментальной (и, со всей очевидностью, непротиворечивой) формальной системы $\mathbb{L}$, в рамках которой выражаются лишь вышеупомянутые простейшие правила логического вывода, — например, с так называемого *исчисления предикатов* (см. [223]), которое только на это и способно, — и построить систему $\mathbb{E}$ посредством присоединения к системе $\mathbb{L}$ процедуры $E$ в виде дополнительных аксиом и правил процедуры для $\mathbb{L}$, переведя тем самым всякое высказывание $P$, получаемое из процедуры $E$, в разряд ИСТИННЫХ. Это, впрочем, вовсе не обязательно окажется легко достижимым на практике. Если процедура $E$ задается всего лишь в виде спецификации машины Тьюринга, то нам, возможно, придется присоединить к системе $\mathbb{L}$ (как часть ее алфавита и правил процедуры) все необходимые обозначения и операции машины Тьюринга, *прежде* чем мы сможем присоединить саму процедуру $E$ в качестве, по сути, дополнительной аксиомы. (См. окончание § 2.8; подробности в [223].)

Собственно говоря, в нашем случае не имеет большого значения, содержит ли система $\mathbb{E}$, которую мы таким образом строим, ИСТИННЫЕ предположения, отличные от тех, что можно получить непосредственно из процедуры $E$ (да и примитивные логические правила системы $\mathbb{L}$ вовсе не обязательно должны являться частью заданной процедуры $E$). В § 2.5 мы рассматривали гипотетический алгоритм $A$, который по определению включал в себя все процедуры (известные или познаваемые), которыми располагают математики для установления факта незавершаемости вычислений. Любому подобному алгоритму неизбежно *придется*, помимо всего прочего, включать в себя и все основные операции простого логического вывода. Поэтому в дальнейшем я буду подразумевать, что все эти вещи в алгоритме $A$ изначально присутствуют.

Следовательно, как процедуры для установления математических истин, алгоритмы (т. е. вычислительные процессы) и формальные системы для нужд моего доказательства, в сущности, *эквивалентны*. Таким образом, несмотря на то, что представленное в § 2.5 доказательство было сформулировано исключительно для вычислений, оно сгодится и для общих формальных систем. В том доказательстве, если помните, речь шла о совокупности всех вычислениях (действий машины Тьюринга) $C_q(n)$. Следовательно, для того чтобы оно оказалось во всех отношениях применимо к формальной системе $\mathbb{F}$, эта система должна быть достаточно обширной для того, чтобы включать в себя действия всех машин Тьюринга. Алгоритмическую процедуру $A$, предназначенную для установления факта незавершаемости некоторых вычислений, мы можем теперь добавить к правилам системы $\mathbb{F}$ с тем, чтобы вычисления, предположения о незавершающемся характере которых устанавливаются в рамках $\mathbb{F}$ как ИСТИННЫЕ, были бы тождественны всем тем вычислениям, незавершаемость которых определяется с помощью процедуры $A$.

Как же первоначальное кенигсбергское доказательство Гёделя связано с тем, что я представил в § 2.5? Не будем углубляться в детали, укажем лишь на наиболее существенные моменты. В роли формальной системы $\mathbb{F}$ из исходной теоремы Гёделя выступает наша алгоритмическая процедура $A$:

$$\text{алгоритм } A \longleftrightarrow \text{правила системы } \mathbb{F}.$$

Роль же представленного Гёделем в Кенигсберге предположения $G(\mathbb{F})$, которое в действительности утверждает непротиворечивость системы $\mathbb{F}$, играет полученное в § 2.5 конкретное предположение «вычисление $C_k(k)$ не завершается», недоказуемое посредством процедуры $A$, но интуитивно представляющееся истинным, коль скоро процедуру $A$ мы полагаем обоснованной:

$$\text{утверждение «вычисление } C_k(k) \text{ не завершается»} \longleftrightarrow \text{утверждение «система } \mathbb{F} \text{ непротиворечива».}$$



Возможно, такая замена позволит лучше понять, каким образом убежденность в обоснованности процедуры — такой, например, как $A$ — может привести к другой процедуре, с исходной никак не связанной, но в обоснованности которой мы *также* должны быть убеждены, поскольку если мы полагаем процедуры некоторой формальной системы $\mathbb{F}$ обоснованными — т. е. процедурами, с помощью которых мы получаем одни лишь действительные математические истины, полностью исключив ложные утверждения (иными словами, если некое предположение $P$ выводится из такой процедуры как ИСТИННОЕ, то это значит, что оно и в самом деле должно *быть истинным*), — то мы должны также уверовать и в $\omega$-непротиворечивость системы $\mathbb{F}$. Если под «ИСТИННЫМ» понимать «истинное», а под «ЛОЖНЫМ» — «ложное» (как оно, собственно, и есть в рамках любой обоснованной формальной системы $\mathbb{F}$), то безусловно истинно следующее утверждение:

не все предположения $P(0)$, $P(1)$, $P(2)$, $P(3)$, $P(4)$, ... могут быть ИСТИННЫМИ, если утверждение «предположение $P(n)$ справедливо для всех натуральных чисел $n$» ЛОЖНО,

что в точности совпадает с условием $\omega$-непротиворечивости.

Однако убежденность в $\omega$-непротиворечивости формальной системы $\mathbb{F}$ может происходить не только из убежденности в обоснованности этой системы, но и из убежденности в ее обыкновенной непротиворечивости. Поскольку если под «ИСТИННЫМ» понимать «истинное», а под «ЛОЖНЫМ» — «ложное», то, несомненно, выполняется условие

«ни одно предположение $P$ не может быть *одновременно* *и* ИСТИННЫМ, *и* ЛОЖНЫМ»,

в точности совпадающее с условием непротиворечивости. Вообще говоря, во многих случаях различия между непротиворечивостью и $\omega$-непротиворечивостью практически отсутствуют. Для

упрощения дальнейших рассуждений этой главы я, в общем случае, не стану разделять эти два типа непротиворечивости и буду обычно говорить просто о «непротиворечивости». Суть доказательства Гёделя и Россера сводится к тому, что установление факта непротиворечивости формальной системы (достаточно обширной) превышает возможности этой самой формальной системы. Первоначальный (кенигсбергский) вариант теоремы Гёделя опирался только на $\omega$-непротиворечивость, однако следующий, более известный, вывод был связан уже исключительно с непротиворечивостью обыкновенной.

Сущность гёделевского доказательства в нашем случае состоит в том, что оно показывает, как выйти за рамки любого заданного набора вычислительных правил, полагаемых обоснованными, и получить некое дополнительное правило, в исходном наборе отсутствующее, но которое также должно полагать обоснованным, — т. е. правило, утверждающее *непротиворечивость* исходных правил. Важно уяснить следующий существенный момент:

убежденность в *обоснованности* равносильна убежденности в *непротиворечивости*.

Мы имеем право применять правила формальной системы $\mathbb{F}$ и полагать, что выводимые из нее результаты действительно *истинны*, только в том случае, если мы также полагаем, что эта формальная система непротиворечива. (Например, если бы система $\mathbb{F}$ не была непротиворечивой, то мы могли бы вывести, как ИСТИННОЕ, утверждение «1 = 2», которое истинным, разумеется, не является!) Таким образом, если мы уверены, что применение правил некоторой формальной системы $\mathbb{F}$ действительно эквивалентно математическому рассуждению, то следует быть готовым принять и рассуждение, выходящее за рамки системы $\mathbb{F}$, *какой бы эта система* $\mathbb{F}$ *ни была*.

## 2.10. Возможные формальные возражения против $\mathscr{G}$ (продолжение)

Продолжим рассмотрение различных математических возражений, высказываемых время от времени в отношении моей трактовки доказательства Гёделя–Тьюринга. Многие из них тесно связаны друг с другом, однако я полагаю, что в любом случае их будет полезно разъяснить по отдельности.



**Q10. Абсолютна ли математическая истина? Как мы уже видели, существуют различные мнения относительно абсолютной истинности утверждений о бесконечных множествах. Можем ли мы доверять доказательствам, опирающимся на какую-то расплывчатую концепцию «математической истины», а не на, скажем, четко определенное понятие формальной ИСТИНЫ?**

Что касается формальной системы $\mathbb{F}$, описывающей общую теорию множеств, то, действительно, не всегда ясно, можно ли вообще говорить о каком-то абсолютном смысле, в котором то или иное утверждение о множествах является либо «истинным», либо «ложным», — вследствие чего под сомнение может попасть и само понятие «обоснованности» формальной системы, подобной $\mathbb{F}$. В качестве поясняющего примера приведем один известный результат, полученный Гёделем (1940) и Коэном (1966). Они показали, что определенные математические утверждения (так называемые *континуум-гипотеза* Кантора и *аксиома выбора*) никак не зависят от теоретико-множественных аксиом системы *Цермело–Френкеля* — стандартной формальной системы, обозначаемой здесь через $\mathbb{ZF}$. (Аксиома выбора гласит, что для любой совокупности непустых множеств существует еще одно множество, которое содержит ровно один элемент из каждого множества совокупности[1]. Согласно же континуум-гипотезе Кантора, количество подмножеств натуральных чисел — равное количеству *вещественных* чисел — представляет собой вторую по величине бесконечность после множества собственно натуральных чисел[2]. Читателю нет нужды вникать в скрытый смысл этих утверждений прямо сейчас. Равно как нет нужды и мне углубляться в подробное изложение аксиом и правил процедуры системы $\mathbb{ZF}$.) Некоторые математики убеждены в том, что система $\mathbb{ZF}$ охватывает все методы математического рассуждения, необходимые для обычной математики. Некоторые даже утверждают, будто приемлемым математическим доказательством можно считать только такое доказательство, какое можно, в принципе, сформулировать и доказать в рамках системы $\mathbb{ZF}$. (См. комментарий к возражению **Q14**, где дается оценка применимости к таким субъектам гёделевского доказательства.) Иными словами, эти математики настаивают на том, что ИС-

ТИННЫМИ, ЛОЖНЫМИ и НЕРАЗРЕШИМЫМИ в рамках системы $\mathbb{ZF}$ математическими утверждениями можно считать только те утверждения, истинность, ложность и неразрешимость которых, в принципе, устанавливается математическими средствами. Для таких людей аксиома выбора и континуум-гипотеза являются математически неразрешимыми (что, по их мнению, и доказывается выводом Гёделя–Коэна), и они наверняка будут утверждать, что истинность или ложность этих двух математических утверждений суть предметы достаточно условные.

Влияют ли эти кажущиеся неопределенности в отношении абсолютного характера математической истины на выводы, которые мы сделали из доказательства Гёделя–Тьюринга? Никоим образом, так как мы имеем здесь дело с классом математических проблем гораздо более ограниченной природы, нежели те, что, подобно аксиоме выбора и континуум-гипотезе, относятся к неконструктивно-бесконечным множествам. В данном случае нас занимают лишь утверждения вида

«такое-то вычисление никогда не завершается»,

причем рассматриваемые вычисления можно задать совершенно точно через действия машины Тьюринга. Такие утверждения в логике называются $\Pi_1$-*высказываниями* (или, точнее, $\Pi_1^0$-высказываниями). В пределах формальной системы $\mathbb{F}$ утверждение $G(\mathbb{F})$ является $\Pi_1$-высказыванием, а вот $\Omega(\mathbb{F})$ таковым не является (см. §2.8). По всей видимости, не существует каких-либо разумных доводов против того, что истинный/ложный характер любого $\Pi_1$-высказывания есть предмет *абсолютный* и никак не зависит от избранного нами мнения относительно предположений, касающихся неконструктивно-бесконечных множеств — таких, например, как аксиома выбора и континуум-гипотеза. (С другой стороны, как мы вскоре убедимся, выбор метода рассуждения, принимаемого нами в качестве инструмента для получения убедительных *доказательств* $\Pi_1$-высказываний, действительно может определяться мнением, которого мы придерживаемся в отношении неконструктивно-бесконечных множеств; см. возражение **Q11**.) Очевидно, если не считать крайней позиции, занимаемой отдельными интуиционистами (см. комментарий к **Q9**), единственное здравое возражение по поводу абсолютного характера истинности таких утверждений может быть связано с тем обстоятельством, что некоторые принципиально



завершающиеся вычисления могут потребовать для своего выполнения столь непомерно долгого времени, что на практике, вполне возможно, не завершатся, скажем, и за все время жизни Вселенной; может случиться и так, что для записи самого вычисления (пусть и конечного) потребуется так много символов, что физически невозможным окажется составить даже его описание. Впрочем, все эти вопросы были исчерпывающим образом проанализированы выше, в обсуждении возражения **Q8**; там же мы выяснили, что на наш основной вывод $\mathscr{G}$ они никоим образом не влияют. Вспомним и о возражении **Q9**, рассмотрение которого показало, что интуиционисты в этом случае также не избегают вывода $\mathscr{G}$.

Кроме того, концепция (весьма ограниченная, надо сказать) математической истины, необходимая мне для доказательства Гёделя–Тьюринга, определена, вообще говоря, не менее четко, нежели концепции ИСТИННОГО, ЛОЖНОГО и НЕРАЗРЕШИМОГО для любой формальной системы $\mathbb{F}$. Из сказанного выше (§2.9) нам известно, что существует некий *алгоритм* $F$, эквивалентный системе $\mathbb{F}$. Если алгоритму $F$ предстоит обработать некое предположение $P$ (формулируемое на языке системы $\mathbb{F}$), то выполнение этого алгоритма может быть успешно завершено только в том случае, если предположение $P$ доказуемо в соответствии с правилами системы $\mathbb{F}$, т. е. когда предположение $P$ ИСТИННО. Соответственно, предположение $P$ является ЛОЖНЫМ, если алгоритм $F$ успешно завершается при обработке предположения $\sim P$, и НЕРАЗРЕШИМЫМ, если не завершается ни одно из упомянутых вычислений. Вопрос о том, является ли математическое утверждение $P$ ИСТИННЫМ, ЛОЖНЫМ или НЕРАЗРЕШИМЫМ, в точности совпадает по своей природе с вопросом о реальной истинности утверждений о завершаемости или незавершаемости вычислений — иными словами, о ложности или истинности определенных $\Pi_1$-высказываний — а кроме этого для нашего «гёделевско–тьюринговского» доказательства ничего и не требуется.

**Q11. Существуют определенные $\Pi_1$-высказывания, которые можно доказать с помощью теории бесконечных множеств, однако не известно ни одного доказательства, которое использовало бы стандартные «конечные» методы. Не означает ли это, что**

**даже к таким четко определенным проблемам математики, на деле, подходят субъективно? Различные математики, придерживающиеся в отношении теории множеств разных убеждений, могут применять к оценке математической истинности $\Pi_1$-высказываний неэквивалентные критерии.**

Этот момент может оказаться существенным в том, что касается моих собственных выводов из доказательства Гёделя (—Тьюринга), и я, возможно, уделил ему недостаточно много внимания в кратком изложении, представленном в НРК. Как ни странно, но возражение **Q11**, похоже, никого, кроме меня, не обеспокоило — по крайней мере, никто мне на него не указал! В НРК (с. 417, 418), как и здесь, я сформулировал доказательство Гёделя(—Тьюринга) исходя из того, что посредством разума и понимания способны установить все «математики» или «математическое сообщество». Преимущество подобной формулировки, в отличие от рассмотрения вопроса о способности какого-либо *конкретного* индивидуума к установлению математических истин посредством своего разума и понимания, заключается в том, что первый способ позволяет избежать некоторых возражений, которые нередко выдвигают в отношении той версии доказательства Гёделя, которую предложил Лукас (1961). Самые разные ученые[3] указывали, к примеру, на то, что «сам Лукас» никак не мог обладать знанием о своем собственном алгоритме. (Некоторые из них говорили то же самое и о варианте доказательства, предложенном мною[4], не обратив, судя по всему, внимания на тот факт, что моя формулировка вовсе не настолько «личностна».) Именно возможность сослаться на способности к рассуждению и пониманию, присущие всем «математикам» вообще или «математическому сообществу», позволяет нам избежать необходимости считаться с предположением о том, что различные индивидуумы могут воспринимать математическую истину поразному, каждый в соответствии с личным непознаваемым алгоритмом. Значительно сложнее смириться с тем, что результатом выполнения некоего непостижимого алгоритма может оказаться коллективное понимание математического сообщества в целом, нежели с тем, что этот самый алгоритм обусловливает математическое понимание всего лишь какого-то конкретного индивидуума. Суть возражения **Q11** как раз и заключается в том, что



упомянутое коллективное понимание может оказаться совсем не таким универсальным и безличным, каким счел его я.

Утверждения, о каких говорится в **Q11**, действительно, существуют. То есть существуют $\Pi_1$-высказывания, единственные известные доказательства которых опираются на то или иное применение теории бесконечных множеств. Такое $\Pi_1$-высказывание может быть результатом арифметического кодирования утверждения типа «аксиомы формальной системы $\mathbb{F}$ являются непротиворечивыми», где система $\mathbb{F}$ подразумевает манипуляции обширными бесконечными множествами, само существование которых может быть сомнительным. Математик, убежденный в реальном *существовании* некоторого достаточно обширного неконструктивного множества **S**, придет к выводу, что система $\mathbb{F}$ действительно непротиворечива, тогда как другой математик, который полагает, что множества **S** не существует, вовсе не обязан считать систему $\mathbb{F}$ непротиворечивой. Таким образом, даже ограничив рассмотрение одним вполне определенным вопросом о завершении или незавершении работы машины Тьюринга (т. е. ложности или истинности $\Pi_1$-высказываний), мы не можем себе позволить не учитывать субъективности *убеждений* в отношении, скажем, существования некоторого обширного неконструктивно-бесконечного множества **S**. Если различные математики используют для установления истинности определенных $\Pi_1$-высказываний *неэквивалентные* «персональные алгоритмы», то, по-видимому, с моей стороны несправедливо говорить о просто «математиках» или «математическом сообществе».

Полагаю, что в строгом смысле это действительно может быть несколько несправедливо; и читатель может при желании перефразировать вывод $\mathscr{G}$ следующим образом:

> $\mathscr{G}^*$ Для установления математической истины ни один отдельно взятый математик не применяет только те алгоритмы, какие он (или она) полагает обоснованными.

Представленные мною доводы по-прежнему остаются в силе, однако, мне кажется, некоторые из более поздних утратят значительную часть своей силы, если представить ситуацию в таком виде. Более того, в случае формулировки $\mathscr{G}^*$ все доказательство уходит в направлении, на мой взгляд, бесперспективном, сосредоточенном, в большей степени, на конкретных механизмах, управляющих действиями конкретных индивидуумов, нежели на

принципах, лежащих в основе действий любого из нас. Меня же на данном этапе интересует не столько различия подходов отдельных математиков к той или иной математической проблеме, сколько то *общее*, что есть между нашим пониманием и нашим математическим восприятием.

Попытаемся разобраться, *действительно* ли мы вынуждены принять формулировку $\mathscr{G}^*$. В самом ли деле суждения математиков настолько субъективны, что они могут *принципиально* расходиться при установлении истинности какого-то конкретного $\Pi_1$-высказывания? (Разумеется, доказательство, устанавливающее истинность $\Pi_1$-высказывания, может быть просто-напросто быть слишком громоздким или слишком сложным, чтобы его мог воспроизвести тот или иной математик (см. ниже по тексту возражение **Q12**), т. е. *на практике* математики вполне могут разойтись во мнениях. Однако в данном случае нас интересует вовсе не это. Мы занимаемся исключительно *принципиальными* вопросами.) Вообще говоря, математическое доказательство есть вещь не настолько субъективная, как может показаться на основании вышесказанного. Математики могут придерживаться самых разных — и, на их взгляд, неопровержимо истинных — точек зрения по тем или иным фундаментальным вопросам и во всеуслышание объявлять об этом, однако едва дело доходит до доказательств или опровержений каких-либо вполне определенных конкретных $\Pi_1$-высказываний, все разногласия тут же куда-то исчезают. Никто не воспримет всерьез доказательство $\Pi_1$-высказывания, утверждающего, по сути своей, непротиворечивость некоторой формальной системы $\mathbb{F}$, если математик будет основывать его только лишь на существовании некоего спорного бесконечного множества $\mathbf{S}$. То, что при этом в действительности доказывается, можно сформулировать следующим, куда более приемлемым, образом: «Если множество $\mathbf{S}$ существует, то формальная система $\mathbb{F}$ является непротиворечивой, и в этом случае данное $\Pi_1$-высказывание истинно».

Тем не менее, могут быть и исключения: например, один математик полагает, что некоторое неконструктивно-бесконечное множество $\mathbf{S}$ «с очевидностью» существует — или, по крайней мере, что допущение о его существовании никоим образом не приводит к противоречию, — другой же математик никакой очевидности здесь не усматривает. Дискуссии математиков по таким *фундаментальным* вопросам могут порой принимать поистине



неразрешимый характер. При этом обе стороны могут оказаться, в принципе, неспособны сколько-нибудь убедительно изложить свои доказательства, даже в отношении $\Pi_1$-высказываний. Возможно, каждому математику и в самом деле присуще некое особое внутреннее восприятие истинности утверждений, связанных с неконструктивно-бесконечными множествами. Конечно же, математики нередко *заявляют* о том, что их восприятие таких вещей в корне отличается от восприятия коллег. Однако я полагаю, что такие различия, по сути своей, подобны различиям в *ожиданиях*, которые различные математики могут иметь и в отношении истинности обычных математических высказываний. Эти ожидания суть всего лишь предварительные предположения. До тех пор, пока не представлено убедительного доказательства или опровержения, математики могут спорить друг с другом об ожидаемой или *предполагаемой* истинности того или иного положения, однако представление такого доказательства одним из математиков убеждает (в принципе) всех. Что до фундаментальных вопросов, то там этих доказательств как раз нет. Возможно, и не будет. Быть может, их нельзя отыскать по той причине, что их просто-напросто нет, а фундаментальные вопросы допускают существование различных, но равно *справедливых* точек зрения.

Здесь, однако, следует подчеркнуть еще один связанный с $\Pi_1$-высказываниями момент. Возможность наличия у математика *ошибочной* точки зрения — т. е. такой точки зрения, которая вынуждает его делать неверные выводы в отношении истинности тех или иных $\Pi_1$-высказываний, — нас в данный момент *не интересует*. Нет ничего невероятного в том, что математики порой опираются на неверное в фактическом отношении «понимание» — а то и на *необоснованные алгоритмы*, — только к настоящему обсуждению это никакого отношения не имеет, поскольку *согласуется* с выводом $\mathscr{G}$. Впрочем, эту ситуацию мы подробно рассмотрим ниже, в § 3.4. Следовательно, дело в данном случае заключается не в том, могут ли разные математики придерживаться *противоречащих* одна другой точек зрения, а скорее в том, может ли одна точка зрения оказаться, в принципе, *мощнее* другой. Каждая такая точка зрения будет совершенно справедлива в том, что касается установления истинности $\Pi_1$-высказываний, однако какая-то из них сможет, в принципе, дать своим последователям возможность установить, что те или иные вычисления не завершаются, тогда как другие, более слабые,

точки зрения на это неспособны; то есть одни математики будут обладать существенно большей способностью к пониманию, нежели другие.

Не думаю, что такая возможность представляет собой сколько-нибудь серьезную угрозу для моей первоначальной формулировки $\mathscr{G}$. Хотя в отношении бесконечных множеств математики и вправе придерживаться различных точек зрения, этих самых точек зрения вовсе не *так* много: по всей видимости, не более пяти. Существенные в этом смысле расхождения могут быть обусловлены лишь утверждениями, подобными аксиоме выбора (о ней говорилось в комментарии к возражению **Q10**), которую одни полагают «очевидной», другие же напрочь отвергают связанную с ней неконструктивность. Любопытно, что эти различные точки зрения на собственно аксиому выбора *не* приводят непосредственно к тому $\Pi_1$-высказыванию, относительно справедливости которого возникают разногласия. Ибо, независимо от своей предполагаемой «истинности» или «ложности», аксиома выбора, как показывает теорема Гёделя—Коэна (см. комментарий к **Q10**), не вступает в противоречие со стандартными аксиомами системы $\mathbb{ZF}$. Могут, однако, существовать и *другие* спорные аксиомы, соответствующей теоремы для которых нет. Впрочем, обыкновенно, когда речь заходит о принятии или опровержении той или иной теоретико-множественной аксиомы — назовем ее аксиомой $Q$, — утверждения математиков принимают следующий вид: «Из допущения справедливости аксиомы $Q$ следует, что... ». Такое утверждение при всем желании не сможет стать предметом спора между математиками. Аксиома выбора, похоже, является исключением в том смысле, что ее справедливость часто подразумевается без приведения упомянутой оговорки, однако это обстоятельство, по-видимому, никак не противоречит моей общей объективной формулировке вывода $\mathscr{G}$ — при условии, что мы ограничимся только $\Pi_1$-высказываниями:

> $\mathscr{G}^{**}$ Для установления истинности $\Pi_1$-высказываний математики-люди не применяют заведомо обоснованные алгоритмы,

а этого нам в любом случае вполне достаточно.

Есть ли другие спорные аксиомы, которые одни математики считают «очевидными», а другие ставят под сомнение? Думаю, будет огромным преувеличением сказать, что имеется хотя



бы десять существенно различных точек зрения на теоретико-множественные допущения, которые в явном виде как допущения не формулируются. Положим, что их не более десяти, и рассмотрим следствия из этого допущения. Это означает, что существует порядка десяти, по сути, различных классов математиков, различаемых по типу рассуждения в отношении бесконечных множеств, который они полагают «очевидно» истинным. Каждого такого математика можно назвать математиком $n$-го класса, где $n$ изменяется в весьма узком диапазоне — не более десяти значений. (Чем больше номер класса, тем мощнее будет точка зрения принадлежащих к нему математиков.) Вывод $\mathscr{G}^{**}$ принимает в этом случае следующий вид:

> $\mathscr{G}^{***}$ Для установления истинности $\Pi_1$-высказываний математики-люди $n$-го класса (где $n$ может принимать лишь несколько значений) не применяют только те алгоритмы, какие они полагают обоснованными.

Так получается, потому что доказательство Гёделя(—Тьюринга) можно применять к каждому классу отдельно. (Важно понять, что само гёделевское доказательство предметом спора между математиками не является, а потому если для любого математика $n$-го класса гипотетический алгоритм $n$-го класса будет познаваемо обоснованным, то доказательство приведет к противоречию.) Таким образом, как и в случае с $\mathscr{G}$, дело вовсе не в существовании какого-то невообразимого количества непознаваемо обоснованных алгоритмов, каждый из которых присущ лишь одному конкретному индивидууму. Мы всего лишь исключаем возможность существования некоторого очень небольшого количества неэквивалентных непознаваемо обоснованных алгоритмов, рассортированных в соответствии с их мощностью и образующих в результате различные «школы мышления». В последующем обсуждении различия между вариантами $\mathscr{G}^{***}$ и $\mathscr{G}$ либо $\mathscr{G}^{**}$ не будут иметь особого значения, поэтому для упрощения изложения я не стану в дальнейшем их как-то различать и буду использовать для них всех одно общее обозначение $\mathscr{G}$.

**Q12. Вне зависимости от того, насколько различных точек зрения придерживаются математики *в принципе*, *на практике* те же математики обладают весьма разными способностями к воспроизведению**

**доказательств, разве не так? Не менее различны и их способности к пониманию, позволяющие им совершать математические открытия.**

Безусловно, так оно и есть, однако к рассматриваемому вопросу все эти вещи не имеют ну абсолютно никакого отношения. Меня не интересует, какие именно и насколько сложные доказательства математик способен воспроизвести *на практике*. Еще меньше меня занимает вопрос о том, какие доказательства математик может на практике *открыть* или какие понимание и вдохновение могут ему в этом способствовать. Здесь мы говорим исключительно о том, доказательства какого типа математики могут, в принципе, воспринимать как обоснованные.

Оговорка «в принципе» используется в наших рассуждениях отнюдь не просто так. Если допустить, что некий математик располагает доказательством или опровержением некоторого $\Pi_1$-высказывания, то его разногласия с другими математиками касательно обоснованности данного доказательства разрешимы только в том случае, если у этих самых других математиков хватит времени, терпения, объективности, способностей и решимости с вниманием и пониманием воспроизвести всю — возможно, длинную и хитроумную — цепочку его рассуждений. На практике же математики вполне могут отказаться от всех этих трудов еще до полного разрешения спорных вопросов. Однако подобные проблемы к данному исследованию отношения не имеют. Так как, по всей видимости, существует все же некий вполне определенный смысл, в котором то, что *в принципе* постижимо для одного математика, оказывается равным образом (если отвлечься на время от возражения **Q11**) постижимо и для другого, — вообще, для любого человека, способного мыслить. Рассуждения бывают весьма громоздкими, а участвующие в них концепции могут показаться чересчур тонкими или туманными, и тем не менее существуют достаточно убедительные основания полагать, что способность к пониманию одного человека не включает в себя ничего такого, что в принципе недоступно другому человеку. Это применимо и к тем случаям, когда для воспроизведения во всех подробностях чисто вычислительной части доказательства может потребоваться помощь компьютера. Возможно, не совсем разумно ожидать, что математик-человек будет лично выполнять все необходимые для такого доказательства вычисления, и все же он, вне всякого



сомнения, сможет без особого труда понять и проверить каждый *отдельный* его этап.

Здесь я говорю исключительно о сложности математического доказательства и ни в коем случае не о возможных существенных и принципиальных вопросах, которые могут вызвать среди математиков разногласия в отношении выбора допустимых методов рассуждения. Разумеется, я встречал математиков, утверждавших, что они в своей практике сталкивались с такими математическими доказательствами, которые были совершенно вне их компетенции: «Я уверен, что, сколько бы я ни старался, мне никогда не понять того-то или такого-то; этот метод рассуждения мне не по зубам». В каждом конкретном случае подобного заявления необходимо индивидуально решать, действительно ли данный метод рассуждения *в принципе* выходит за рамки системы убеждений этого математика — каковой случай мы рассматривали в комментарии к возражению **Q11**, — или он вообще-то *смог бы* разобраться в принципах, на которых основано это доказательство, если бы только приложил больше сил и затратил больше времени. Как правило, справедливым оказывается последнее. Более того, источником отчаяния нашего математика чаще всего становится туманный стиль изложения или ограниченные лекторские способности «такого-то», а вовсе не то, что какие-то существенные и принципиальные моменты «того-то» действительно выходят за рамки его способностей. Толковое изложение, на первый взгляд, непонятного предмета чудесным образом устраняет все прежние недоразумения.

Чтобы еще раз подчеркнуть, что я имею в виду, скажу следующее: сам я часто посещаю математические семинары, на которых не слежу (а иногда и не пытаюсь следить) за подробностями представляемых доказательств. Наверное, если бы я сел где-нибудь и обстоятельно изучил эти самые доказательства, я и в самом деле смог бы проследить за мыслью автора — хотя, возможно, это удалось бы мне лишь при наличии дополнительной литературы или устных пояснений, которые восполнили бы возможные пробелы в моем образовании или же в материалах самого семинара. Я знаю, что в действительности я этого делать не стану. У меня почти наверняка не окажется на это ни времени, ни достаточного количества внимания, ни, впрочем, особого желания. Но при этом я вполне могу принять представленный на семинаре результат на веру по всевозможным «несуществен-

ным» причинам — например, потому что полученный результат правдоподобно «выглядит», или потому что у лектора надежная репутация, или потому что другие слушатели, которых я считаю более сведущими в таких делах, нежели я сам, этот результат оспаривать не стали. Конечно, я могу ошибиться во всех своих умозаключениях, а результат вполне может оказаться ложным — либо истинным, но никоим образом не следующим из представленного доказательства. Все эти тонкости никак не влияют на ту принципиальную позицию, которую я здесь представляю. Результат может оказаться истинным и адекватно доказанным, и в таком случае я, *в принципе*, могу проследить за ходом всего доказательства — или же ошибочным, в каковом случае, как уже упоминалось, он нас в данном контексте не интересует (см. § 3.2 и § 3.4). Возможные исключения могут составить лишь те случаи, когда представляемый материал касается каких-либо спорных аспектов теории бесконечных множеств или опирается на какой-то необычный метод рассуждения, который может быть признан сомнительным в соответствии с теми или иными математическими воззрениями (что, само по себе, может заинтриговать меня до такой степени, что я впоследствии действительно попытаюсь это доказательство повторить). Как раз такие исключительные ситуации мы обсуждали выше, в комментарии к возражению **Q11**.

Что касается подобных соображений относительно природы математической точки зрения, на практике многие математики могут и не иметь четкого представления о том, каких именно фундаментальных принципов они в действительности придерживаются. Однако, как уже было сказано выше, в комментарии к **Q11**, если математик, у которого нет определенной позиции в отношении того, следует ли принимать, скажем, некую «аксиому $Q$», желает проявить осмотрительность, то ничто не мешает ему изложить требующие принятия аксиомы $Q$ результаты в следующем виде: «Из принятия аксиомы $Q$ следует, что...». Разумеется, математики, несмотря на всю их пресловутую педантичность, проявляют в подобных вопросах должную осмотрительность далеко не всегда. Нельзя отрицать и того, что время от времени им удается допускать и вовсе очевидные ошибки. И все же все эти ошибки — если они допущены по недосмотру, а не следуют из тех или иных непоколебимых принципов — являются *исправимыми*. (Как упоминалось ранее, возможность действительного применения математиками в качестве основы для своих решений необос-



нованного алгоритма будет подробно рассмотрена в § 3.2 и § 3.4. Поскольку эта возможность *не противоречит* выводу $\mathscr{G}$, она не является предметом настоящего обсуждения.) В данном случае нас не занимают исправимые ошибки, так как к вопросу о принципиальной достижимости тех или иных результатов они никакого отношения не имеют. А вот возможные неопределенности в действительных взглядах математиков, безусловно, требуют дальнейшего обсуждения, которое и приводится ниже.

**Q13. У математиков нет *абсолютно* определенных убеждений относительно обоснованности или непротиворечивости используемых ими формальных систем — как нет и однозначного ответа на вопрос о том, «пользователями» *каких* именно формальных систем они себя полагают. Не подвергаются ли их убеждения постепенному размыванию по мере того, как формальные системы все более удаляются от области феноменов, доступных непосредственному интуитивному или экспериментальному восприятию?**

И правда, нечасто встретишь математика, способного похвалиться прочно устоявшимися и непоколебимо непротиворечивыми убеждениями, когда речь заходит об основах предмета. Кроме того, по мере накопления опыта математик вполне может изменить свои взгляды относительно того, что считать неопровержимо истинным, если он вообще склонен считать неопровержимо истинным *что бы то ни было*. Можно ли, например, быть совершенно и полностью уверенным в том, что число 1 отлично от числа 2? Если говорить о некоей *абсолютной* человеческой уверенности, то не совсем ясно, можно ли подобное понятие как-то однозначно определить. Однако какую-то точку опоры все же выбрать необходимо. Вполне приемлемой точкой опоры может стать принятие в качестве неопровержимо истинной *некоторой* системы убеждений и принципов, от которой уже можно двигаться в своих рассуждениях дальше. Разумеется, нельзя забывать и о том, что многие математики вовсе не имеют определенного мнения относительно того, что именно можно считать неопровержимо истинным. Таких математиков я попросил бы какую-никакую опору для себя все же выбрать и просто быть готовыми при необходимости впоследствии ее сменить. Как показывает

доказательство Гёделя, *какую бы* позицию математик в этом случае ни занял, ее все равно невозможно полностью уместить в рамки правил любой постижимой формальной системы (а если и возможно, то этот факт невозможно однозначно установить). И дело даже не в том, что та или иная конкретная позиция постоянно изменяется; система убеждений, полностью охватываемая рамками *любой* (достаточно обширной) формальной системы $\mathbb{F}$, неизбежно должна также простирается и за пределы доступной $\mathbb{F}$ области. Любая позиция, среди неопровержимых убеждений которой имеется и убеждение в обоснованности системы $\mathbb{F}$, должна также включать в себя и убежденность в истинности гёделевского предположения[7] $G(\mathbb{F})$. Убежденность в истинности $G(\mathbb{F})$ не представляет собой изменения позиции; эта убежденность уже подразумевается неявно в исходной позиции, допускающей принятие истинности формальной системы $\mathbb{F}$, пусть даже поначалу это и не очевидно.

Безусловно, всегда существует возможность того, что в выводы, получаемые математиком на основании исходных посылок какой-либо конкретной точки зрения, закрадется ошибка. Одна только *возможность* возникновения такой ошибки — даже если в действительности никакой ошибки допущено не было — может привести к уменьшению степени убежденности, которую математик питает в отношении своих выводов. Однако такое «постепенное размывание» нас, вообще говоря, не занимает. Подобно действительным ошибкам, оно «исправимо». Более того, если доказательство было проведено действительно корректно, то чем дольше его изучаешь, тем, как правило, более убедительными представляются полученные в нем выводы. «Постепенное размывание» математик может испытать *на практике*, но не в принципе, что возвращает нас к обсуждению возражения **Q12**.

Таким образом, вопрос перед нами встает здесь следующий: имеет ли место постепенное размывание *в принципе*, т. е. может ли математик счесть, скажем, обоснованность некоторой формальной системы $\mathbb{F}$ неопровержимой, тогда как в обоснованности более сильной системы $\mathbb{F}^*$ он будет лишь «практически уверен». Этот вопрос не представляется мне сколько-нибудь серьезным, коль скоро, какой бы ни была система $\mathbb{F}$, мы вправе настаивать, чтобы она включала в себя обычные логические правила и арифметические операции. Упомянутый выше математик, который верит в обоснованность системы $\mathbb{F}$, должен также верить в ее непротиворечивость, а следовательно, и в истинность гёделевского высказывания $G(\mathbb{F})$. Таким образом, одни только выводы из формальной системы $\mathbb{F}$ не могут охватывать всей совокупности математических убеждений математика, *какой бы* эта система ни была.

[7] Пояснение к используемым здесь обозначениям можно найти в § 2.8. Впрочем, $G(\mathbb{F})$ без ущерба для смысла рассуждения можно было бы везде заменить на $\Omega(\mathbb{F})$, в чем мы убедимся ниже.



Однако следует ли считать высказывание $G(\mathbb{F})$ *неопровержимо* истинным всякий раз, когда мы признаем неопровержимо обоснованной формальную систему $\mathbb{F}$? Полагаю, утвердительный ответ на этот вопрос не должен вызывать никаких сомнений; и это тем более так, если придерживаться в отношении воспроизведения математического доказательства той «принципиальной» позиции, которой мы придерживались до сих пор. Единственная возникающая в этой связи реальная проблема касается деталей фактического кодирования утверждения «система $\mathbb{F}$ непротиворечива» в форме арифметического утверждения ($\Pi_1$-высказывания). Сама по себе базовая *идея* неопровержимо очевидна: если система $\mathbb{F}$ является обоснованной, то она, безусловно, непротиворечива. (Так как если бы она не была непротиворечивой, то среди ее утверждений присутствовало бы утверждение «$1 = 2$», т. е. система была бы необоснованной.) Что касается деталей этого самого кодирования, то здесь нам вновь предстоит иметь дело с различием между «принципиальным» и «практическим» уровнями. Не составит особого труда убедиться в том, что такое кодирование в принципе возможно (хотя сам процесс убеждения может занять некоторое время), однако убедиться в корректном выполнении того или иного конкретного *действительного* кодирования — дело совсем другое. Детали кодирования, как правило, бывают в известной степени произвольными и в разных изложениях могут весьма значительно отличаться. Возможно, где-то закрадется незначительная ошибка или просто опечатка, которая, в формальном смысле, должна бы сделать недействительным данное конкретное предназначенное для выражения «$G(\mathbb{F})$» теоретико-числовое предположение, однако в действительности этого не происходит.

Надеюсь, читатель понимает, что возможность возникновения таких ошибок не существенна, когда речь заходит о том, что мы подразумеваем здесь под принятием предположения $G(\mathbb{F})$ в

качестве неопровержимой истины. Я, разумеется, говорю о *действительном* предположении $G(\mathbb{F})$, а не о возможном случайном предположении, непреднамеренно сформулированном благодаря опечатке или незначительной ошибке. В этой связи мне вспоминается одна история о великом американском физике Ричарде Фейнмане. Фейнман, по-видимому, объяснял одному из студентов какое-то понятие, но оговорился. Когда студент выразил недоумение, Фейнман вспылил: «Не слушайте, что я говорю; слушайте, что я *имею в виду*!»[8].

Один из возможных способов такого явного кодирования состоит в использовании представленных еще в НРК спецификаций машин Тьюринга и точном воспроизведении доказательства гёделевского типа, описанного в §2.5 (пример такого кодирования приводится в Приложении А). Впрочем, даже и в этом случае об абсолютной «явности» говорить нельзя, поскольку нам понадобится еще и каким-то явным образом закодировать правила формальной системы $\mathbb{F}$ в системе обозначений действий машин Тьюринга; обозначим такой код, скажем, через $T_{\mathbb{F}}$. (Код $T_{\mathbb{F}}$ должен удовлетворять определенному свойству: если некоторому высказыванию $P$, выводимому в рамках системы $\mathbb{F}$, ставится в соответствие некоторое число $p$, то необходимо, скажем, чтобы равенство $T_{\mathbb{F}}(p) = 1$ выполнялось всякий раз, когда высказывание $P$ является теоремой системы $\mathbb{F}$, в противном же случае вычисление $T_{\mathbb{F}}(p)$ не должно завершаться вовсе.) Безусловно, все это открывает широкий простор для формальных ошибок. Помимо возможных трудностей, связанных с практическим построением кода $T_{\mathbb{F}}$ на основе системы $\mathbb{F}$ и отысканием числа $p$ на основе высказывания $P$, отсутствует ясность и в отношении другого вопроса: а не ошибся ли я сам где-нибудь в спецификациях машин Тьюринга, — иными словами, можем ли мы быть полностью уверены в корректности приведенного в Приложении А этой книги кода, если вдруг решим использовать для отыскания вычисления $C_k(k)$ именно это определение? Лично я думаю, что ошибок там нет, однако в собственной непогрешимости я уверен куда как меньше, нежели в первоначальных построениях Гёделя (пусть и более сложных). Впрочем, всякому дочитавшему до этого места, смею надеяться, уже ясно, что возможные ошибки подобного рода существенной роли здесь не играют. Помните, что говорил Фейнман?

[8] Источник цитаты мне, к сожалению, обнаружить не удалось. Однако, как справедливо заметил Рихард Йожа, точная формулировка слов Фейнмана не имеет никакого значения, поскольку послание, которое они несут, применимо и к ним самим!



Что же касается собственно моих спецификаций, следует упомянуть еще один формальный момент. Представленный мною в §2.5 вариант доказательства Гёделя(–Тьюринга) опирается не на непротиворечивость системы $\mathbb{F}$, а на обоснованность алгоритма $A$, и являет собой критерий для установления незавершаемости вычислений (т. е. истинности $\Pi_1$-высказываний). Этот вариант подходит нам ничуть не хуже любых других, поскольку известно, что из обоснованности алгоритма $A$ следует истинность утверждения о незавершаемости вычисления $C_k(k)$, каковое явное утверждение (тоже $\Pi_1$-высказывание) мы имеем полное право использовать вместо высказывания $G(\mathbb{F})$. Более того, как отмечалось выше (см. §2.8), доказательство, вообще говоря, зависит не от непротиворечивости формальной системы $\mathbb{F}$, а от ее $\omega$-непротиворечивости. Из обоснованности системы $\mathbb{F}$ очевидно следует ее непротиворечивость, равно как и $\omega$-непротиворечивость. Если допустить, что система $\mathbb{F}$ обоснованна, то ни $\Omega(\mathbb{F})$, ни $G(\mathbb{F})$ из ее правил (см. §2.8) не следуют, однако оба эти высказывания являются истинными.

Думаю, можно с уверенностью заключить, что какое бы «постепенное размывание» убежденности того или иного математика ни сопровождало переход от убеждения в обоснованности формальной системы $\mathbb{F}$ к убеждению в истинности высказывания $G(\mathbb{F})$ (или $\Omega(\mathbb{F})$), оно будет целиком и полностью обусловлено возможностью ошибки в точной формулировке полученного им высказывания «$G(\mathbb{F})$». (То же применимо и к высказыванию $\Omega(\mathbb{F})$.) Все это не имеет непосредственного отношения к настоящему обсуждению — при наличии *подлинной* (не случайной) формулировки высказывания $G(\mathbb{F})$ никакого размывания убежденности происходить не должно. Если формальная система $\mathbb{F}$ неопровержимо обоснованна, то *ее* высказывание $G(\mathbb{F})$ столь же неопровержимо истинно. Все формы заключения $\mathcal{G}$ ($\mathcal{G}^{**}$, $\mathcal{G}^{***}$) остаются неизменными при условии, что под «истинностью» подразумевается «неопровержимая истинность».

**Q14. Нет никаких сомнений в том, что формальная система $\mathbb{ZF}$ — или некоторая стандартная ее модификация (обозначим ее через $\mathbb{ZF}^*$) — действитель-**

**но включает в себя все необходимое для серьезной математической деятельности. Почему бы просто не принять эту систему за основу, смириться с недоказуемостью ее непротиворечивости и продолжить свои математические изыскания?**

Полагаю, такая точка зрения весьма и весьма распространена среди практикующих математиков, особенно тех, кто не слишком углубляется в фундаментальные основы или философию своего предмета. Подобное отношение вполне естественно для людей, главной заботой которых является просто хорошее выполнение серьезной, пусть и математической, работы (хотя *в действительности* такие люди крайне редко выражают свои результаты в рамках строгих правил формальных систем, подобных $\mathbb{ZF}$). Согласно этой точке зрения, математика имеет дело лишь с тем, что можно доказать или опровергнуть в рамках некоей конкретной формальной системы — такой, например, как $\mathbb{ZF}$ (или какая-либо ее модификация $\mathbb{ZF}^*$). С высоты такой позиции математическая деятельность и в самом деле напоминает своего рода «игру». Назовем ее $\mathbb{ZF}$-*игрой* (или $\mathbb{ZF}^*$-игрой), причем играть в эту игру следует в соответствии с правилами, установленными в рамках данной системы. Такой подход характерен для *формалиста*, подлинный же формалист мыслит исключительно в терминах ИСТИННОГО и ЛОЖНОГО, которые не обязательно совпадают с истинным и ложным в их повседневном смысле. Если формальная система обоснованна, то все, что является ИСТИННЫМ, и будет истинным, а все, что ЛОЖНО, будет ложным. Однако наверняка найдутся высказывания, формализуемые в рамках данной системы, которые, будучи истинными, не являются ИСТИННЫМИ, и другие, которые, будучи ложными, не являются ЛОЖНЫМИ, иными словами, в обоих случаях эти высказывания оказываются НЕРАЗРЕШИМЫМИ. Если система $\mathbb{ZF}$ непротиворечива, то в $\mathbb{ZF}$-игре гёделевское высказывание[9] $G(\mathbb{ZF})$ и его отрицание $\sim G(\mathbb{ZF})$ принадлежат, соответственно, к этим двум категориям. (Более того, окажись система $\mathbb{ZF}$ противоречивой, то *и* высказывание $G(\mathbb{ZF})$, *и* его отрицание $\sim G(\mathbb{ZF})$ были бы ИСТИННЫМИ и ЛОЖНЫМИ одновременно!)

[9] Как и ранее, обозначение $G(\mathbb{F})$ можно без каких бы то ни было последствий заменить на $\Omega(\mathbb{F})$. То же справедливо и для комментариев к **Q15**–**Q20**.



$\mathbb{ZF}$-игра, судя по всему, представляет собой исключительно разумный подход, позволяющий реализовать большую часть того, что нас интересует в обычной математике. Однако по причинам, которые обозначены выше, я совершенно не в состоянии понять, каким же образом из нее может «произрасти» реальная точка зрения в отношении чьих бы то ни было математических *убеждений*. Ибо если кто-то считает, что с помощью «практикуемой» им математики он устанавливает исключительно подлинные математические истины — скажем, истинность $\Pi_1$-высказываний, — то он должен верить и в то, что используемая им система *обоснованна*; а если он верит в ее обоснованность, то он должен также верить в ее *непротиворечивость*, то есть в то, что $\Pi_1$-высказывание, утверждающее истинность $G(\mathbb{F})$, *действительно* истинно, несмотря на то, что оно НЕРАЗРЕШИМО. Таким образом, математические убеждения человека должны включать в себя нечто, что в рамках $\mathbb{ZF}$-игры невыводимо. С другой стороны, если человек не верит в обоснованность формальной системы $\mathbb{ZF}$, то он не может верить и в подлинную истинность ИСТИННЫХ результатов, полученных с помощью $\mathbb{ZF}$-игры. В обоих случаях сама по себе $\mathbb{ZF}$-игра не в состоянии снабдить нас удовлетворительной позицией в том, что касается математической истинности. (Это равным образом применимо к любой формальной системе $\mathbb{ZF}^*$.)

**Q15. Выбранная нами формальная система $\mathbb{F}$ может и *не* оказаться непротиворечивой — по крайней мере, мы не можем быть вполне *уверены* в ее непротиворечивости; по какому же, в таком случае, праву мы утверждаем, что высказывание $G(\mathbb{F})$ «очевидно» истинно?**

Хотя этот вопрос был достаточно исчерпывающе рассмотрен в предыдущих обсуждениях, я полагаю, что суть того рассмотрения полезно будет изложить еще раз, поскольку возражения, подобные **Q15**, чаще всего оказываются среди нападок на наше с Лукасом приложение теоремы Гёделя. Суть же в том, что мы вовсе не утверждаем, что высказывание $G(\mathbb{F})$ непременно истинно для любой формальной системы $\mathbb{F}$, мы утверждаем лишь, что высказывание $G(\mathbb{F})$ настолько же достоверно, насколько достоверна любая другая истина, получаемая применением правил

самой системы $\mathbb{F}$. (Вообще говоря, высказывание $G(\mathbb{F})$ оказывается *более* достоверным, нежели утверждения, получаемые действительным *применением* правил $\mathbb{F}$, так как система $\mathbb{F}$, даже *будучи* непротиворечивой, не обязательно будет обоснованной!) Если мы верим в истинность любого утверждения $P$, выводимого исключительно с помощью правил системы $\mathbb{F}$, то мы должны верить и в истинность $G(\mathbb{F})$, по крайней мере, в той же степени, в какой мы верим в истинность $P$. Таким образом, ни одна постижимая формальная система $\mathbb{F}$ — или эквивалентный ей алгоритм $F$ — не может послужить абсолютно полной основой для подлинного математического познания или формирования убеждений. Как отмечалось в комментариях к **Q5** и **Q6**, наше доказательство построено как *reductio ad absurdum*: мы выдвигаем предположение, что система $\mathbb{F}$ действительно является абсолютной основой для формирования убеждений, а затем показываем, что такое предположение приводит к противоречию, т. е. является неверным.

Мы, конечно же, можем, как в **Q14**, выбрать для удобства какую-то конкретную систему $\mathbb{F}$, хотя уверенности в том, что она обоснованна, а потому непротиворечива, это нам не добавит. Впрочем, при наличии *действительных* сомнений в обоснованности системы $\mathbb{F}$ любой получаемый в рамках $\mathbb{F}$ результат $P$ следует формулировать в виде

«высказывание $P$ выводимо в рамках системы $\mathbb{F}$»

(или, что то же самое, «высказывание $P$ ИСТИННО»), избегая утверждений вида «высказывание $P$ истинно». Такое утверждение в математическом смысле вполне приемлемо и может быть либо действительно истинным, либо действительно ложным. Совершенно законным образом мы можем свести все наши математические высказывания к утверждениям такого рода, однако и в этом случае нам никуда не деться от утверждений об абсолютных математических истинах. При случае мы можем прийти к убеждению, будто мы установили, что какое-то утверждение вышеприведенного вида является в действительности ложным, т. е. получить следующий результат:

«высказывание $P$ невыводимо в рамках системы $\mathbb{F}$».

Такие утверждения имеют вид: «такое-то вычисление не завершается» (или, по сути, «будучи примененным к высказыванию $P$,



алгоритм $F$ не завершается»), что в точности совпадает с формой рассматриваемых нами $\Pi_1$-высказываний. Вопрос: какие средства мы полагаем допустимыми в процессе получения подобных утверждений? Каковы, наконец, те математические процедуры, в которые мы действительно *верим* и применяем при установлении математических истин? Такая система убеждений, при условии, что они достаточно разумны, никак не может быть *эквивалентна* всего лишь убежденности в обоснованности и непротиворечивости формальной системы, какой бы эта формальная система ни была.

**Q16. Заключение об истинности высказывания $G(\mathbb{F})$ для непротиворечивой формальной системы $\mathbb{F}$ мы делаем, исходя из допущения, что те символы системы $\mathbb{F}$, которые, как мы полагаем, служат для представления натуральных чисел, *действительно* представляют натуральные числа. Окажись на их месте другие числа — скажем, некие экзотические «сверхнатуральные» числа, — мы вполне могли бы обнаружить, что высказывание $G(\mathbb{F})$ ложно. Откуда мы знаем, что в нашей системе $\mathbb{F}$ мы имеем дело с натуральными, а не со «сверхнатуральными» числами?**

В самом деле, конечного аксиоматического способа убедиться в том, что «числа», о которых идет речь, и есть те самые подразумеваемые *натуральные* числа, а не какие-то посторонние «сверхнатуральные», не существует[5]. Однако, в некотором смысле, в этом и состоит вся суть гёделевского рассуждения. Неважно, какую именно схему аксиом формальной системы $\mathbb{F}$ мы построим, пытаясь охарактеризовать натуральные числа, — одних лишь правил системы $\mathbb{F}$ будет недостаточно, чтобы определить, является ли высказывание $G(\mathbb{F})$ действительно истинным или же ложным. Полагая систему $\mathbb{F}$ непротиворечивой, мы знаем, что в высказывании $G(\mathbb{F})$ *подразумевается* все же наличие некоего истинного смысла. Это, однако, происходит лишь в том случае, если символы, составляющие в действительности формальное выражение, обозначаемое «$G(\mathbb{F})$», имеют подразумеваемые значения. Если эти символы интерпретировать как-либо иначе, то полученная в результате интерпретация «$G(\mathbb{F})$» вполне может оказаться ложной.

Для того чтобы разобраться, откуда берутся все эти двусмысленности, рассмотрим новые формальные системы $\mathbb{F}^*$ и $\mathbb{F}^{**}$, где $\mathbb{F}^*$ получается путем присоединения к аксиомам системы $\mathbb{F}$ высказывания $G(\mathbb{F})$, а $\mathbb{F}^{**}$ — путем аналогичного присоединения высказывания $\sim G(\mathbb{F})$. Если система $\mathbb{F}$ обоснованна, то обе системы $\mathbb{F}^*$ и $\mathbb{F}^{**}$ непротиворечивы (т. к. высказывание $G(\mathbb{F})$ истинно, а $\sim G(\mathbb{F})$ из правил системы $\mathbb{F}$ вывести невозможно. При этом в случае подразумеваемой (или *стандартной*) интерпретации символов $\mathbb{F}$ из обоснованности системы $\mathbb{F}$ следует, что система $\mathbb{F}^*$ обоснованна, а система $\mathbb{F}^{**}$ — *нет*. Впрочем, одним из характерных свойств непротиворечивых формальных систем является возможность отыскания так называемых *нестандартных* реинтерпретаций символов таким образом, что высказывания, которые являются ложными в стандартной интерпретации, оказываются истинными в нестандартной; соответственно, в такой нестандартной интерпретации обоснованными могут быть системы $\mathbb{F}$ и $\mathbb{F}^{**}$, а система $\mathbb{F}^*$ обоснованной не будет. Можно вообразить, что такая реинтерпретация может повлиять на смысл логических символов (таких как «$\sim$» и «&», которые в стандартной интерпретации означают, соответственно, «не» и «и»), однако в данном случае нас занимают символы, обозначающие неопределенные числа («$x$», «$y$», «$z$», «$x'$», «$x''$» и т. д.), и значения применяемых к ним логических кванторов ($\forall$, $\exists$). В стандартной интерпретации символы «$\forall x$» и «$\exists x$» означают, соответственно, «для всех натуральных чисел $x$» и «существует такое натуральное число $x$, что»; в *не*стандартной же интерпретации эти символы могут относится не к натуральным числам, а к числам какого-то иного вида с иными свойствами упорядочения (такие числа действительно можно назвать «сверхнатуральными», или даже «ультранатуральными», как это сделал Хофштадтер [201]).

Дело, однако, в том, что мы-то *знаем*, что такое на самом деле представляют собой натуральные числа, и для нас не составит никакого труда отличить их от каких-то непонятных сверхнатуральных чисел. Натуральные числа суть самые обыденные вещи, обозначаемые, как правило, символами 0, 1, 2, 3, 4, 5, 6, .... С этой концепцией мы знакомимся еще в детском возрасте и легко отличим ее от надуманной концепции сверхнатурального числа (см. § 1.21). Есть что-то таинственное в том, что мы, похоже, и впрямь обладаем каким-то инстинктивным пониманием действительного смысла понятия натурального числа. Все, что



мы получаем в этом смысле в детском (или уже взрослом) возрасте, сводится к сравнительно небольшому количеству описаний понятий «нуля», «единицы», «двух», «трех» и т. д. («три апельсина», «один банан» и т. п.), однако при этом, несмотря на всю неадекватность такого описания, мы как-то умудряемся постичь всю концепцию в целом. В некотором платоническом смысле натуральные числа видятся своего рода категориями, обладающими абсолютным концептуальным существованием, от нас никак не зависящим. И все же, несмотря на «человеконезависимость» натуральных чисел, мы оказываемся способны установить интеллектуальную связь с действительной концепцией натуральных чисел, опираясь лишь на неоднозначные и, на первый взгляд, неадекватные описания. С другой стороны, не существует конечного набора *аксиом*, с помощью которого можно было бы провести четкую границу между множеством натуральных чисел и альтернативным ему множеством так называемых «сверхнатуральных» чисел.

Более того, такое специфическое свойство всей совокупности натуральных чисел, как их *бесконечное* количество, мы также можем каким-то образом воспринимать непосредственно, тогда как система, действие которой ограничено точными конечными правилами, не способна отличить данную конкретную бесконечность натуральных чисел от других возможных («сверхнатуральных») вариантов. Мы же легко понимаем бесконечность, характеризующую натуральные числа, пусть и обозначаем ее просто точками «...» —

«0, 1, 2, 3, 4, 5, 6, ...»,

либо сокращением «и т. д.» —

«нуль, один, два, три и т. д.».

Нам не нужно объяснять на языке каких-то точных правил, что именно представляет собой натуральное число. В этом смысле можно считать, что нам повезло, так как такое объяснение дать невозможно. Как только нам приблизительно укажут верное направление, мы тут же обнаруживаем, что уже откуда-то *знаем*, что это за штука такая — натуральное число!

Возможно, некоторые читатели знакомы с *аксиомами Пеано* для арифметики натуральных чисел (об арифметике Пеано я уже упоминал в § 2.7), и, возможно, теперь эти читатели находятся в некотором недоумении: почему же аксиомы Пеано не дают

адекватного определения натуральных чисел. Согласно определению Пеано, мы начинаем ряд натуральных чисел с символа **0** и затем добавляем слева особый «оператор следования», обозначаемый **S** и осуществляющий простое прибавление единицы к числу, над которым совершается действие, т. е. **1** *определяется* как **S0**, **2** как **S1** или **SS0** и т. д. В качестве правил мы располагаем следующими утверждениями: если **Sa=Sb**, то **a=b**; и ни при каком **x** число **0** нельзя записать в виде **Sx** (последнее утверждение служит для характеристики числа **0**). Кроме того, имеется «принцип индукции», согласно которому некое свойство чисел (скажем, $P$) должно быть истинным в отношении *всех* чисел **n**, если оно удовлетворяет двум условиям: (i) если истинно $P$ (**n**), то для всех **n** истинно также и $P$ (**Sn**); (ii) $P$ (**0**) истинно. Сложности начинаются, когда дело доходит до логических операций, символы которых $\forall$ и $\exists$ в стандартной интерпретации означают, соответственно, «для всех *натуральных чисел* ...» и «существует такое *натуральное число* ..., что». В нестандартной интерпретации смысл этих символов соответствующим образом изменяется, так что они квантифицируют уже не натуральные числа, а «числа» какого-то другого типа. Хотя математические спецификации Пеано, задающие оператор следования **S**, действительно описывают отношение упорядочения, отличающее натуральные числа от разных прочих «сверхнатуральных» чисел, эти определения невозможно записать в терминах формальных правил, которым удовлетворяют кванторы $\forall$ и $\exists$. Для того чтобы передать смысл математических определений Пеано, необходимо перейти к так называемой «логике второго порядка», в которой также вводятся кванторы типа $\forall$ и $\exists$, но только теперь они оперируют не над отдельными натуральными числами, а над *множествами* (бесконечными) натуральных чисел. В «логике первого порядка» арифметики Пеано кванторы оперируют над отдельными числами, и в результате получается формальная система в обычном смысле этого слова. Логика же второго порядка нам формальной системы не дает. В случае строгой формальной системы вопрос о правильности применения правил системы решается чисто *механическими* (т. е. алгоритмическими) способами — в сущности, именно это свойство формальных систем и послужило причиной их рассмотрения в настоящем контексте. В рамках логики второго порядка упомянутое свойство не работает.



Многие ошибочно полагают (в духе приведенных в возражении **Q16** соображений), что из теоремы Гёделя следует существование множества различных арифметик, каждая из которых в равной степени обоснованна. Соответственно, та частная арифметика, которую мы, возможно, по чистой случайности избрали для своих нужд, определяется просто какой-то произвольно взятой формальной системой. В действительности же теорема Гёделя показывает, что ни одна из этих формальных систем (будучи непротиворечивой) не может быть полной; поэтому (как доказывается далее) к ней можно непрерывно добавлять какие угодно новые аксиомы и получать всевозможные альтернативные непротиворечивые системы, которыми при желании можно заменить ту, в рамках которой мы работаем в настоящий момент. Эту ситуацию нередко сравнивают с той, что сложилась некогда с евклидовой геометрией. На протяжении двадцати одного века люди верили, что евклидова геометрия является единственно возможной геометрией. Но когда в восемнадцатом веке сразу несколько великих математиков (таких как Гаусс, Лобачевский и Бойяи) показали, что существуют в равной степени возможные альтернативы общепринятой геометрии, геометрии пришлось отступить с абсолютных позиций на произвольные. Нередко можно услышать, будто Гёдель показал, что арифметика так же представляет собой предмет произвольного выбора, при этом один набор непротиворечивых аксиом оказывается ничуть не хуже любого другого.

Однако подобная интерпретация того, что доказал Гёдель, абсолютно неверна. Согласно Гёделю, само по себе понятие формальной системы аксиом не подходит для передачи даже самых элементарных математических понятий. Когда мы употребляем термин «арифметика» без дальнейших пояснений, мы подразумеваем обычную арифметику, которая работает с обычными натуральными числами 0, 1, 2, 3, 4, ... (и, быть может, с их отрицаниями), а вовсе не со «сверхнатуральными» числами, что бы это понятие ни означало. Мы можем, если пожелаем, исследовать свойства формальных систем, и это, конечно же, станет ценным вкладом в процесс математического познания. Однако такое предприятие несколько отличается от исследования обычных свойств обычных натуральных чисел. В некотором отношении данная ситуация весьма напоминает ту, что сложилась в последнее время с геометрией. Изучение неевклидовых геометрий

интересно с математической точки зрения, да и сами геометрии имеют ряд важных областей применения (например, в физике, см. НРК, глава 5, особенно рис. 5.1 и 5.2, а также § 4.4), но, когда термин «геометрия» используется в обычном языке (в отличие от «жаргона» математиков или физиков-теоретиков), подразумевается, как правило, обычная евклидова геометрия. Однако имеется и разница: то, что логик может назвать «евклидовой геометрией», действительно можно определить (с некоторыми оговорками[6]) через определенную формальную систему, тогда как обычную «арифметику», как показал Гёдель, определить таким образом нельзя.

Гёдель доказал не то, что математика (в особенности арифметика) — это произвольные поиски, направление которых определяется прихотью Человека; он доказал, что математика — это нечто абсолютное, и в ней мы должны не изобретать, но открывать (см. § 1.17). Мы открываем, что такое натуральные числа и без труда отличаем их от любых сверхнатуральных чисел. Гёдель показал, что ни одна система «искусственных» правил не способна сделать это за нас. Такая платоническая точка зрения была существенна для Гёделя, не менее существенной она будет и для нас в последующих рассуждениях (§ 8.7).

**Q17. Допустим, что формальная система $\mathbb{F}$ предназначена для представления тех математических истин, что в принципе доступны человеческому разуму. Не можем ли мы обойти проблему невозможности формального включения в систему $\mathbb{F}$ гёделевского высказывания $G(\mathbb{F})$, включив вместо него что-либо, имеющее смысл $G(\mathbb{F})$, воспользовавшись при этом новой интерпретацией смысла символов системы $\mathbb{F}$?**

Определенные способы представления примененного к $\mathbb{F}$ гёделевского доказательства в рамках формальной системы $\mathbb{F}$ (достаточно обширной) действительно существуют, коль скоро новый, реинтерпретированный, смысл символов системы $\mathbb{F}$ полагается отличным от исходного смысла символов этой системы. Однако если мы пытаемся таким образом интерпретировать систему $\mathbb{F}$ *как* процедуру, с помощью которой разум приходит к тем или иным математическим выводам, то подобный подход является не чем иным, как шулерством. Если мы намерены толковать



мыслительную деятельность исключительно в рамках системы $\mathbb{F}$, то ее символы не должны изменять свой смысл «на полпути». Если же мы принимаем, что мыслительная деятельность может содержать что-то помимо операций самой системы $\mathbb{F}$ — т. е. изменение *смысла* символов, — то нам необходимо знать и правила, управляющие подробным изменением. Либо эти правила окажутся неалгоритмическими, и это сыграет в пользу $\mathscr{G}$, либо для них найдется какая-то конкретная алгоритмическая процедура, и тогда нам следовало бы изначально включить *эту* процедуру в нашу «систему $\mathbb{F}$» — обозначим ее через $\mathbb{F}^\dagger$ — с тем, чтобы она представляла собой полную совокупность процедур, обусловливающих наши с вами понимание и проницательность, а значит, необходимости в изменении смысла символов не возникло бы вовсе. В последнем случае вместо гёделевского высказывания $G(\mathbb{F})$ из предыдущего рассуждения нам предстоит разбираться уже с высказыванием $G(\mathbb{F}^\dagger)$, так что ничего мы в результате не выигрываем.

**Q18. Даже в такой простой системе, как арифметика Пеано, можно сформулировать теорему, интерпретация которой имеет следующий смысл:**

**«система $\mathbb{F}$ обоснованна», а следовательно, «высказывание $G(\mathbb{F})$ истинно».**

**Разве это не все, что нам нужно от теоремы Гёделя? Значит, теперь, полагая обоснованной какую угодно формальную систему $\mathbb{F}$, мы вполне можем поверить и в истинность ее гёделевского высказывания — при условии, разумеется, что мы готовы принять арифметику Пеано, разве не так?**

Подобную теорему[7] действительно можно сформулировать в рамках арифметики Пеано. Точнее (поскольку мы не можем в пределах какой бы то ни было формальной системы должным образом выразить понятие «обоснованности» или «истинности», как это следует из знаменитой теоремы Тарского), мы, в сущности, формулируем более сильный результат:

«система $\mathbb{F}$ непротиворечива», а следовательно, «высказывание $G(\mathbb{F})$ истинно»,

либо иначе:

«система $\mathbb{F}$ $\omega$-непротиворечива», а следовательно, «высказывание $\Omega(\mathbb{F})$ истинно».

Из этих высказываний следует вывод, необходимый для **Q18**, поскольку если система $\mathbb{F}$ обоснованна, то она, разумеется, непротиворечива или омега-непротиворечива, в зависимости от обстоятельств. Понимая *смысл* присутствующего здесь символизма, мы и в самом деле можем поверить в истинность высказывания $G(\mathbb{F})$ на основании одной лишь веры в обоснованность системы $\mathbb{F}$. Это, впрочем, мы уже приняли. Если понимать смысл, то действительно возможно перейти от $\mathbb{F}$ к $G(\mathbb{F})$. Сложности возникнут лишь в том случае, если нам вздумается исключить необходимость интерпретаций и сделать переход от $\mathbb{F}$ к $G(\mathbb{F})$ *автоматическим*. Будь это возможно, мы смогли бы автоматизировать общую процедуру «гёделизации» и создать алгоритмическое устройство, которое действительно будет содержать в себе все, что нам нужно от теоремы Гёделя. Однако такой возможности у нас нет — захоти мы добавить эту предполагаемую алгоритмическую процедуру в какую угодно формальную систему $\mathbb{F}$, выбранную нами в качестве отправной, в результате просто-напросто получилась бы, по сути, некоторая *новая* формальная система $\mathbb{F}^\sharp$, а *ее* гёделевское высказывание $G(\mathbb{F}^\sharp)$ оказалось бы уже за ее рамками. Таким образом, согласно теореме Гёделя, *какой-то* аспект понимания всегда остается «за нами», независимо от того, какая доля его оказалась включена в формализованную или алгоритмическую процедуру. Это «гёделево понимание» требует постоянного соотнесения с действительным смыслом символов какой бы то ни было формальной системы, к которой применяется процедура Гёделя. В этом смысле ошибка **Q18** весьма похожа на ту, что мы обнаружили, комментируя возражение **Q17**. С невозможностью автоматизации процедуры гёделизации тесно связаны также рассуждения по поводу **Q6** и **Q19**.

В возражении **Q18** присутствует еще один аспект, который стóит рассмотреть. Представим себе, что у нас есть обоснованная формальная система $\mathbb{H}$, содержащая арифметику Пеано. Теорема, о которой говорилось в **Q18**, окажется среди следствий системы $\mathbb{H}$, а частным ее примером, применимым к конкретной системе $\mathbb{F}$ (т. е., собственно, $\mathbb{H}$), будет теорема системы $\mathbb{H}$. Таким образом, можно сформулировать один из выводов формальной системы $\mathbb{H}$:

«система $\mathbb{H}$ обоснованна», а следовательно,
«высказывание $G(\mathbb{H})$ истинно»;



или, точнее, скажем так:

«система $\mathbb{H}$ непротиворечива», а следовательно,
«высказывание $G(\mathbb{H})$ истинно».

Если говорить о реальном смысле этих утверждений, то из них, в сущности, следует, что высказывание $G(\mathbb{H})$ также утверждается системой. А так как (что касается первого из двух вышеприведенных утверждений) истинность *любого* производимого системой $\mathbb{H}$ утверждения, во всяком случае, обусловлена допущением, что система $\mathbb{H}$ обоснованна, то получается, что если система $\mathbb{H}$ утверждает нечто, явно обусловленное ее собственной обоснованностью, то она вполне может утверждать это напрямую. (Из утверждения «если мне можно верить, то $X$ истинно» следует более простое утверждение, исходящее из того же источника: «$X$ истинно».) Однако в действительности обоснованная формальная система $\mathbb{H}$ *не может* утверждать истинность высказывания $G(\mathbb{H})$, что является следствием ее неспособности утверждать собственную обоснованность. Более того, как мы видим, она не может включать в себя и смысл символов, которыми оперирует. Те же факты годятся и для иллюстрации второго утверждения, причем в этом случае ко всему прочему добавляется и некоторая ирония: система $\mathbb{H}$ не способна утверждать собственную непротиворечивость лишь в том случае, если она *действительно* непротиворечива, если же формальная система непротиворечивой *не* является, то подобные ограничения ей неведомы. Противоречивая формальная система $\mathbb{H}$ может утверждать (в качестве «теоремы») вообще все, что она в состоянии сформулировать! Она вполне может, как выясняется, сформулировать и утверждение: «система $\mathbb{H}$ непротиворечива». Формальная система (достаточно обширная) утверждает собственную непротиворечивость тогда и только тогда, когда она *противоречива*!

**Q19. Почему бы нам просто не учредить процедуру многократного добавления высказывания $G(\mathbb{F})$ к любой системе $\mathbb{F}$, какой мы в данным момент пользуемся, и не позволить этой процедуре выполняться *бесконечно*?**

Когда нам дана какая-либо конкретная формальная система $\mathbb{F}$, достаточно обширная и полагаемая обоснованной, мы в состоянии понять, как добавить к ней высказывание $G(\mathbb{F})$ в качестве новой аксиомы и получить тем самым новую систему $\mathbb{F}_1$,

которая также будет считаться обоснованной. (Для согласования обозначений в последующем изложении систему $\mathbb{F}$ можно также обозначить через $\mathbb{F}_0$.) Теперь мы можем добавить к системе $\mathbb{F}_1$ высказывание $G(\mathbb{F}_1)$, получив в результате новую систему $\mathbb{F}_2$, также, предположительно, обоснованную. Повторив данную процедуру, т. е. добавив к системе $\mathbb{F}_2$ высказывание $G(\mathbb{F}_2)$, получим систему $\mathbb{F}_3$ и т. д. Приложив еще совсем немного усилий, мы непременно сообразим, как построить еще одну формальную систему $\mathbb{F}_\omega$, аксиомы которой позволят нам включить в систему в качестве дополнительных аксиом для $\mathbb{F}$ *все* бесконечное множество высказываний $\{G(\mathbb{F}_0), G(\mathbb{F}_1), G(\mathbb{F}_2), G(\mathbb{F}_3), \ldots\}$. Очевидно, что система $\mathbb{F}_\omega$ также будет обоснованной. Этот процесс можно продолжить и дальше: к системе $\mathbb{F}_\omega$ добавляется высказывание $G(\mathbb{F}_\omega)$, в результате чего получается система $\mathbb{F}_{\omega+1}$, к которой затем добавляется высказывание $G(\mathbb{F}_{\omega+1})$, что дает систему $\mathbb{F}_{\omega+2}$, и т. д. Далее, как и в предыдущий раз, мы можем построить формальную систему $\mathbb{F}_{\omega 2}(=\mathbb{F}_{\omega+\omega})$, включив в нее *весь* бесконечный набор соответствующих аксиом, каковая система опять-таки окажется очевидно обоснованной. Добавлением к ней высказывания $G(\mathbb{F}_{\omega 2})$, получим систему $\mathbb{F}_{\omega 2+1}$ и т. д., а потом построим новую систему $\mathbb{F}_{\omega 3}(=\mathbb{F}_{\omega 2+\omega})$, включив в нее опять-таки бесконечное множество аксиом. Повторив всю вышеописанную процедуру, мы сможем получить формальную систему $\mathbb{F}_{\omega 4}$, после следующего повтора — систему $\mathbb{F}_{\omega 5}$ и т. д. Еще чуть-чуть потрудиться, и мы обязательно увидим, как можно включить уже *это* множество новых аксиом $\{G(\mathbb{F}_\omega), G(\mathbb{F}_{\omega 2}), G(\mathbb{F}_{\omega 3}), G(\mathbb{F}_{\omega 4}), \ldots\}$ в новую формальную систему $\mathbb{F}_{\omega^2}(=\mathbb{F}_{\omega\omega})$. Повторив всю процедуру, мы получим новую систему $\mathbb{F}_{\omega^2+\omega^2}$, затем — систему $\mathbb{F}_{\omega^2+\omega^2+\omega^2}$ и т. д.; в конце концов, когда мы сообразим, как связать *все это* вместе (разумеется, и на этот раз не без некоторого напряжения умственных способностей), наши старания приведут нас к еще более всеобъемлющей системе $\mathbb{F}_{\omega^3}$, которая также должна быть обоснованной.

Читатели, которые знакомы с понятием канторовых *трансфинитных ординалов*, несомненно, узнают индексы, обычно используемые для обозначения таких чисел. Тем же, кто от подобных вещей далек, не стóит беспокоиться из-за незнания точного значения этих символов. Достаточно сказать, что описанную процедуру «гёделизации» можно продолжить и далее: мы получим формальные системы $\mathbb{F}_{\omega^4}$, $\mathbb{F}_{\omega^5}$, ..., после чего придем



к еще более обширной системе $\mathbb{F}_{\omega^\omega}$, затем процесс продолжается до еще бóльших ординалов, например, $\omega^{\omega^\omega}$ и т. д. — до тех пор, пока мы все еще способны на каждом последующем этапе понять, каким образом систематизировать все множество гёделизаций, которые мы получили на данный момент. В этом и заключается основная проблема: для упомянутых нами «усилий, трудов и напряжений» требуется соответствующее понимание того, как дóлжно систематизировать предыдущие гёделизации. Эта систематизация выполнима при условии, что достигаемый к каждому последующему моменту этап будет помечаться так называемым *рекурсивным* ординалом, что, в сущности, означает, что должен существовать определенный алгоритм, способный такую процедуру генерировать. Однако алгоритмической процедуры, которую можно было бы заложить заранее и которая позволила бы выполнить описанную систематизацию для *всех* рекурсивных ординалов раз и навсегда, просто-напросто не существует. Нам снова неизбежно потребуется понимание.

Вышеприведенная процедура была впервые предложена Аланом Тьюрингом в его докторской диссертации (а опубликована в [368])[8]; там же Тьюринг показал, что *любое* истинное $\Pi_1$-высказывание можно, в некотором смысле, доказать с помощью многократной гёделизации, подобной описанной нами. (См. также [117].) Впрочем, воспользоваться этим для получения механической процедуры установления истинности $\Pi_1$-высказываний нам не удастся по той простой причине, что механически систематизировать гёделизацию невозможно. Более того, невозможность «автоматизации» процедуры гёделизации как раз и *выводится* из результата Тьюринга. А в §2.5 мы уже показали, что общее установление истинности (либо ложности) $\Pi_1$-высказываний невозможно произвести с помощью *каких бы то ни было* алгоритмических процедур. Так что в поисках систематической процедуры, не доступной тем вычислительным соображениям, которые мы рассматривали до настоящего момента, многократная гёделизация нам ничем помочь не сможет. Таким образом, для вывода $\mathscr{G}$ возражение **Q19** угрозы не представляет.

**Q20. Реальная ценность математического понимания состоит, безусловно, не в том, что благодаря ему мы способны выполнять невычислимые действия,**

**а в том, что оно позволяет нам заменить невероятно сложные вычисления сравнительно простым пониманием. Иными словами, разве не правда, что, используя разум, мы, скорее, «срезаем углы» в смысле теории сложности, а вовсе не «выскакиваем» за пределы вычислимого?**

Я вполне готов поверить в то, что *на практике* интуиция математика гораздо чаще используется для «обхода» вычислительной сложности, чем невычислимости. Как-никак математики по природе своей склонны к лени, а потому зачастую стараются изыскать всяческие способы избежать вычислений (пусть даже им придется в итоге выполнить значительно более сложную мыслительную работу, нежели потребовало бы собственно вычисление). Часто случается так, что попытки заставить компьютеры бездумно штамповать теоремы даже умеренно сложных формальных систем быстро загоняют эти самые компьютеры в ловушку фактически безнадежной вычислительной сложности, тогда как математик-человек, вооруженный пониманием смысла, лежащего в основе правил такой системы, без особого труда получит в рамках этой системы множество интересных результатов(9).

Причина того, что в своих доказательствах я рассматривал не сложность, а невычислимость, заключается в том, что только с помощью последней мне удалось сформулировать необходимые для доказательства сильные утверждения. Не исключено, что в работе большинства математиков вопросы невычислимости играют весьма незначительную роль, если вообще играют. Однако суть не в этом. Я глубоко убежден, что понимание (в частности, математическое) представляет собой нечто, недоступное вычислению, а одной из немногих возможностей вообще подступиться ко всем этим вопросам является как раз доказательство Гёделя (—Тьюринга). Никто не отрицает, что наши математические интуиция и понимание нередко используются для получения результатов, *достижимых*, в принципе, и вычислительным путем, — но и здесь слепое, не отягощенное пониманием, вычисление может оказаться неэффективным настолько, что попросту не будет работать (см. § 3.26). Однако рассмотрение всех таких случаев представляется мне неизмеримо более сложным подходом, нежели обращение к общей невычислимости.



Как бы то ни было, высказанные в возражении **Q20** соображения, пусть и справедливые, все же ни в коей мере не противоречат выводу $\mathscr{G}$.

## Примечания

1. Кому-то, возможно, покажется, что это совершенно «очевидно» и уж никак не может служить предметом спора среди математиков! Проблема, однако, существует, и возникает она в связи с понятием «существования» применительно к большим бесконечным множествам. (См., например, [350], [329], [266].) На примере парадокса Рассела мы уже убедились, что в таких вопросах необходимо проявлять особую осторожность.

   Согласно одной точке зрения, множество не считается необходимо существующим, если нет четкого *правила* (не обязательно вычислимого), устанавливающего, какие элементы в это множество следует включать, а какие — нет. Как раз этого правила аксиома выбора нам и *не* предоставляет, поскольку в ней нет правила, определяющего, *какой* элемент следует взять из каждого множества совокупности. (Некоторые из следствий аксиомы выбора интуитивно не понятны и почти парадоксальны. Вероятно, в этом и состоит одна из причин возникновения разногласий по данному вопросу. Более того, я не совсем уверен, что знаю, какой позиции придерживаюсь в этом отношении *я сам*!)

2. В заключительной главе своей книги, написанной в 1966 году, Коэн подчеркивает, что, хотя он и показал, что континуум-гипотеза является НЕРАЗРЕШИМОЙ в рамках процедур системы $\mathbb{ZF}$, вопрос о том, является ли она действительно *истинной*, был оставлен им без внимания, — и выдвигает некоторые предположения относительно того, каким образом этот вопрос можно действительно *решить*! То есть Коэн, со всей очевидностью, *не* считает, что выбор между принятием или непринятием континуум-гипотезы есть предмет абсолютно произвольный. Это расходится с нередко высказываемым относительно следствий из результатов Гёделя—Коэна мнением, суть которого сводится к тому, что существуют многочисленные «альтернативные теории множеств», для математики в равной степени «справедливые». Такие замечания свидетельствуют о том, что Коэн, подобно Гёделю, является подлинным платонистом, для которого вопросы математической истины ни в коем случае не произвольны, но *абсолютны*. Очень похожих взглядов придерживаюсь и я, см. § 8.7.

3. См., например, [202], [37].

4. См., например, различные комментарии, приведенные в *Behavioral and Brain Sciences*, **13** (1990), 643–705.
5. Терминология была предложена Хофштадтером в [202]. Согласно «другой» теореме Гёделя — так называемой теореме о *полноте*, — подобные нестандартные модели существуют всегда.
6. Вообще говоря, это зависит от того, какие именно утверждения считать частью так называемой «евклидовой геометрии». Если пользоваться обычной терминологией логиков, то система «евклидовой геометрии» включает только утверждения некоторого частного вида, причем оказывается, что истинность или ложность этих утверждений можно определить с помощью алгоритмической процедуры; отсюда и утверждение, что евклидову геометрию можно описать с помощью формальной системы. Однако *в других* интерпретациях обычная «арифметика» тоже могла бы считаться частью «евклидовой геометрии», что допустило бы классы утверждений, которые *невозможно* разрешить алгоритмическим путем. То же самое произошло бы, если бы мы рассмотрели задачу о замощении плоскости полиомино как составляющую евклидовой геометрии, что, казалось бы, вполне естественно. В этом смысле описать геометрию Евклида формально ничуть не проще, чем арифметику!
7. См. комментарий М. Дэвиса в [74].
8. См. также [231], [232] и [163].
9. О некоторых проблемах, с которыми сталкивались компьютерные системы, пытавшиеся самостоятельно «делать математику», можно прочесть у Д. Фридмана [124]. Отметим, что в общем случае такие системы не слишком преуспели. Они по-прежнему остро нуждаются в помощи человека.

# ПРИЛОЖЕНИЕ А: ГЁДЕЛИЗИРУЮЩАЯ МАШИНА ТЬЮРИНГА В ЯВНОМ ВИДЕ

Допустим, что у нас имеется некая алгоритмическая процедура $A$, которая, как нам известно, корректно устанавливает незавершаемость тех или иных вычислений. Мы получим вполне явную процедуру для построения на основе процедуры $A$ конкретного вычисления $C$, для которого $A$ оказывается неадекватной; при этом мы сможем убедиться, что вычисление $C$ действительно *не* завершается. Приняв это явное выражение для $C$, мы сможем определить степень его сложности и сравнить ее со сложностью процедуры $A$, чего требуют аргументы § 2.6 (возражение **Q8**) и § 3.20.

Для определенности я воспользуюсь спецификациями той конкретной машины Тьюринга, которую я описал в НРК. Подробное описание этих спецификаций читатель сможет найти в названной работе. Здесь же я дам лишь краткое описание, которого вполне должно хватить для наших настоящих целей.

Машина Тьюринга имеет конечное число внутренних состояний, но производит все операции на бесконечной ленте. Эта лента представляет собой линейную последовательность «ячеек», причем каждая ячейка может быть маркированной или пустой, а общее количество отметок на ленте — величина конечная. Обозначим каждую маркированную ячейку символом 1, а каждую пустую ячейку — 0. В машине Тьюринга имеется также считывающее устройство, которое поочередно рассматривает отметки и, в явной зависимости от внутреннего состояния машины Тьюринга и характера рассматриваемой в данный момент отметки, определяет дальнейшие действия машины по следующим трем пунктам: (i) следует ли изменить рассматриваемую в данный момент отметку; (ii) каким будет новое внутреннее состояние машины; (iii) должно ли устройство сдвинуться по ленте на один

шаг вправо (обозначим это действие через **R**) или влево (обозначим через **L**), или же на один шаг вправо с остановкой машины (**STOP**). Когда машина, в конце концов, остановится, на ленте слева от считывающего устройства будет представлен в виде последовательности символов 0 и 1 ответ на выполненное ею вычисление. Изначально лента должна быть абсолютно чистой, за исключением отметок, описывающих исходные данные (в виде конечной строки символов 1 и 0), над которыми машина и будет выполнять свои операции. Считывающее устройство в начале работы располагается слева от всех отметок.

При представлении на ленте натуральных чисел (будь то входные или выходные данные) иногда удобнее использовать так называемую *расширенную двоичную* запись, согласно которой число, в сущности, записывается в обычной двоичной системе счисления, только двоичный знак «1» представляется символами 10, а двоичный знак «0» — символом 0. Таким образом, мы получаем следующую схему перевода десятичных чисел в расширенные двоичные:

$$
\begin{aligned}
0 &\longleftrightarrow 0\\
1 &\longleftrightarrow 10\\
2 &\longleftrightarrow 100\\
3 &\longleftrightarrow 1010\\
4 &\longleftrightarrow 1000\\
5 &\longleftrightarrow 10010\\
6 &\longleftrightarrow 10100\\
7 &\longleftrightarrow 101010\\
8 &\longleftrightarrow 10000\\
9 &\longleftrightarrow 100010\\
10 &\longleftrightarrow 100100\\
11 &\longleftrightarrow 1001010\\
12 &\longleftrightarrow 101000\\
13 &\longleftrightarrow 1010010\\
14 &\longleftrightarrow 1010100\\
15 &\longleftrightarrow 10101010\\
16 &\longleftrightarrow 100000\\
17 &\longleftrightarrow 1000010
\end{aligned}
$$

и т. д.



Заметим, что в расширенной двоичной записи символы 1 никогда не встречаются рядом. Таким образом, последовательность из двух или более 1 вполне может послужить сигналом о начале и конце записи натурального числа. То есть для записи всевозможных команд на ленте мы можем использовать последовательности типа 110, 1110, 11110 и т. д.

Отметки на ленте также можно использовать для спецификации конкретных машин Тьюринга. Это необходимо, когда мы рассматриваем работу *универсальной* машины Тьюринга $U$. Универсальная машина $U$ работает с лентой, начальная часть которой содержит подробную спецификацию некоторой конкретной машины Тьюринга $T$, которую универсальной машине предстоит смоделировать. Данные, с которыми должна работать сама машина $T$, подаются в $U$ вслед за тем участком ленты, который определяет машину $T$. Для спецификации машины $T$ можно использовать последовательности 110, 1110 и 11110, которые будут обозначать, соответственно, различные команды для считывающего устройства машины $T$, например: переместиться по ленте на один шаг вправо, на один шаг влево, либо остановиться, сдвинувшись на один шаг вправо:

$$
\begin{aligned}
\mathbf{R} &\longleftrightarrow 110\\
\mathbf{L} &\longleftrightarrow 1110\\
\mathbf{STOP} &\longleftrightarrow 11110.
\end{aligned}
$$

Каждой такой команде предшествует либо символ 0, либо последовательность 10, что означает, что считывающее устройство должно пометить ленту, соответственно, либо символом 0, либо 1, заменив тот символ, который оно только что считало. Непосредственно перед вышеупомянутыми 0 или 10 располагается расширенное двоичное выражение числа, описывающего следующее внутреннее состояние, в которое должна перейти машина Тьюринга согласно этой самой команде. (Отметим, что внутренние состояния, поскольку количество их конечно, можно обозначать последовательными натуральными числами 0, 1, 2, 3, 4, 5, 6, ..., $N$. При кодировании на ленте для обозначения этих чисел будет использоваться расширенная двоичная запись.)

Конкретная команда, к которой относится данная операция, определяется внутренним состоянием машины перед нача-

лом считывания ленты и собственно символами 0 или 1, которые наше устройство при следующем шаге считает и, возможно, изменит. Например, частью описания машины $T$ может оказаться команда 230 → 171R, что означает следующее: «Если машина $T$ находится во внутреннем состоянии 23, а считывающее устройство встречает на ленте символ 0, то его следует заменить символом 1, перейти во внутреннее состояние 17 и переместиться по ленте на один шаг вправо». В этом случае часть «171R» данной команды будет кодироваться последовательностью 100001010110. Разбив ее на участки 1000010.10.110, мы видим, что первый из них представляет собой расширенную двоичную запись числа 17, второй кодирует отметку 1 на ленте, а третий — команду «переместиться на шаг вправо». А как нам описать предыдущее внутреннее состояние (в данном случае 23) и считываемую в соответствующий момент отметку на ленте (в данном случае 0)? При желании можно задать их так же явно с помощью расширенной двоичной записи. Однако на самом деле в этом нет необходимости, поскольку для этого будет достаточно упорядочить различные команды в виде цифровой последовательности (например, такой: 00 → , 01 → , 10 → , 11 → , 20 → , 21 → , 30 → , ...).

К этому, в сущности, и сводится все кодирование машин Тьюринга, предложенное в НРК, однако для завершенности картины необходимо добавить еще несколько пунктов. Прежде всего, следует проследить за тем, чтобы каждому внутреннему состоянию, действующему на отметки 0 и 1 (не забывая, впрочем, о том, что команда для внутреннего состояния с наибольшим номером, действующая на 1, оказывается необходимой не всегда), была сопоставлена какая-либо команда. Если та или иная команда вообще не используется в программе, то необходимо заменить ее «пустышкой». Предположим, например, что в ходе выполнения программы внутреннему состоянию 23 нигде не придется сталкиваться с отметкой 1 — соответствующая команда-пустышка в этом случае может иметь следующий вид: 231 → 00R.

Согласно вышеприведенным предписаниям, в кодированной спецификации машины Тьюринга на ленте пара символов 00 должна быть представлена последовательностью 00, однако можно поступить более экономно и записать просто 0, что явится ничуть не менее однозначным разделителем двух последовательностей, составленных из более чем одного символа 1 подряд[10]. Машина Тьюринга начинает работу, находясь во внутреннем состоянии 0; считывающее устройство движется по ленте, сохраняя это внутреннее состояние до тех пор, пока не встретит первый символ 1. Это обусловлено допущением, что в набор команд машины Тьюринга всегда входит операция 00 → 00R. Таким образом, в действительной спецификации машины Тьюринга в виде последовательности 0 и 1 явного задания этой команды не требуется; вместо этого мы начнем с команды 01 → X, где X обозначает первую нетривиальную операцию запущенной машины, т. е. первый символ 1, встретившийся ей на ленте. Это значит, что начальную последовательность 110 (команду → 00R), которая в противном случае непременно присутствовала бы в определяющей машину Тьюринга последовательности, можно спокойно удалить. Более того, в такой спецификации мы будем всегда удалять и завершающую последовательность 110, так как она одинакова для всех машин Тьюринга.



Получаемая в результате последовательность символов 0 и 1 представляет собой самую обыкновенную (т. е. нерасширенную) *двоичную запись номера машины Тьюринга* $n$ для данной машины (см. главу 2 НРК). Мы называем ее $n$-й машиной Тьюринга и обозначаем $T = T_n$. Каждый такой двоичный номер (с добавлением в конце последовательности 110) есть последовательность символов 0 и 1, в которой нигде не встречается более четырех 1 подряд. Номер $n$, не удовлетворяющий данному условию, определяет «фиктивную машину Тьюринга», которая

[10] Это означает, что при кодировании машины Тьюринга каждую последовательность ...110011... можно заменить на ...11011.... В спецификации универсальной машины Тьюринга, описанной в НРК (см. примечание 7 после главы 2), имеется пятнадцать мест, где я этого не сделал. Чрезвычайно досадная оплошность с моей стороны, и это после того, как я приложил столько усилий, чтобы добиться (в рамках моих же собственных правил) по возможности наименьшего номера, определяющего эту универсальную машину. Упомянутая простая замена позволяет уменьшить мой номер более чем в 30 000 раз! Я благодарен Стивену Ганхаусу за то, что он указал мне на этот недосмотр, а также за то, что он самостоятельно проверил всю представленную в НРК спецификацию и подтвердил, что она *действительно* определяет универсальную машину Тьюринга.

прекратит работать, как только встретит «команду», содержащую более четырех 1. Такую машину «$T_n$» мы будем называть *некорректно определенной*. Ее работа с *какой угодно* лентой является *по определению* незавершающейся. Аналогично, если действующая машина Тьюринга встретит команду перехода в состояние, определенное числом, бóльшим всех тех чисел, для которых были явно заданы возможные последующие действия, то она также «зависнет»: такую машину мы будем полагать «фиктивной», а ее работу — незавершающейся. (Всех этих неудобств можно без особого трýда избежать с помощью тех или иных технических средств, однако реальной необходимости в этом нет; см. § 2.6, **Q4**).

Для того чтобы понять, как на основе заданного алгоритма $A$ построить явное незавершающееся вычисление, факт незавершаемости которого посредством алгоритма $A$ установить невозможно, необходимо предположить, что алгоритм $A$ задан в виде машины Тьюринга. Эта машина работает с лентой, на которой кодируются два натуральных числа $p$ и $q$. Мы полагаем, что если завершается вычисление $A(p, q)$, то вычисление, производимое машиной $T_p$ с числом $q$, *не* завершается вовсе. Вспомним, что если машина $T_p$ определена некорректно, то ее работа с числом $q$ не завершается, каким бы это самое $q$ ни было. В случае такого «запрещенного» $p$ исход вычисления $A(p, q)$ может, согласно исходным допущениям, быть каким угодно. Соответственно, нас будут интересовать исключительно те числа $p$, для которых машина $T_p$ определена *корректно*. Таким образом, в записанном на ленте двоичном выражении числа $p$ пяти символов 1 подряд содержаться не может. Значит, для обозначения на ленте начала и конца числа $p$ мы вполне можем воспользоваться последовательностью 11111.

То же самое, очевидно, необходимо сделать и для числа $q$, причем оно вовсе *не* обязательно должно быть числом того же типа, что и $p$. Здесь перед нами возникает техническая проблема, связанная с чрезвычайной громоздкостью машинных предписаний в том виде, в каком они представлены в НРК. Удобным решением этой проблемы может стать запись чисел $p$ и $q$ в *пятеричной* системе счисления. (В этой системе запись «10» означает число *пять*, «100» — *двадцать пять*, «44» — *двадцать четыре* и т. д.) Однако вместо пятеричных цифр 0, 1, 2, 3 и 4 я воспользуюсь соответствующими последовательностями символов на ленте 0, 10, 110, 1110 и 11110. Таким образом, мы будем записывать



| | | |
|---|---|---|
| 0 | как | 0 |
| 1 | // | 10 |
| 2 | // | 110 |
| 3 | // | 1110 |
| 4 | // | 11110 |
| 5 | // | 100 |
| 6 | // | 1010 |
| 7 | // | 10110 |
| 8 | // | 101110 |
| 9 | // | 1011110 |
| 10 | // | 1100 |
| 11 | // | 11010 |
| 12 | // | 110110 |
| 13 | // | 1101110 |
| 14 | // | 11011110 |
| 15 | // | 11100 |
| 16 | // | 111010 |
| ... | | ... |
| 25 | // | 1000 |
| 26 | // | 10010 |

и т. д.

Под «$C_p$» здесь будет пониматься вычисление, выполняемое корректно определенной машиной Тьюринга $T_r$, где $r$ есть число, обыкновенное двоичное выражение которого (с добавлением в конце последовательности символов 110) в точности совпадает с числом $p$ в нашей пятеричной записи. Число $q$, над которым производится вычисление $C_p$, также необходимо представлять в пятеричном выражении. Вычисление же $A(p, q)$ задается в виде машины Тьюринга, выполняющей действие с лентой, на которой кодируется пара чисел $p$, $q$. Запись на ленте будет выглядеть следующим образом:

...00111110p111110q11111000...,

где **p** и **q** суть вышеописанные пятеричные выражения чисел, соответственно, $p$ и $q$.

Требуется отыскать такие числа $p$ и $q$, для которых не завершается не только вычисление $C_p\,(q)$, но и вычисление $A(p,\,q)$. Процедура из § 2.5 позволяет сделать это посредством отыскания такого числа $k$, при котором вычисление $C_k$, производимое с числом $n$, в точности совпадает с вычислением $A(n,\,n)$ при любом $n$, и подстановки $p = q = k$. Для того чтобы проделать это же в явном виде, отыщем машинное предписание $K\ (= C_k)$, действие которого на последовательность символов на ленте

...**00111110n11111000**...

(где **n** есть пятеричная запись числа $n$) в точности совпадает с действием алгоритма $A$ на последовательность

...**00111110n111110n11111000**...

при любом $n$. Таким образом, действие предписания $K$ сводится к тому, чтобы взять число $n$ (записанное в пятеричном выражении) и однократно его скопировать, при этом два **n** разделяются последовательностью **111110** (та же последовательность начинает и завершает всю последовательность отметок на ленте). Следовательно, оно воздействует на получаемую в результате ленту точно так, как на эту же ленту воздействовал бы алгоритм $A$.

Явную модификацию алгоритма $A$, дающую такое предписание $K$, можно произвести следующим образом. Сначала находим в определении $A$ начальную команду **01 → X** и отмечаем для себя, что это в действительности за «**X**». Мы подставим это выражение вместо «**X**» в спецификации, представленной ниже. Один технический момент: следует, помимо прочего, положить, чтобы алгоритм $A$ был составлен таким образом, чтобы машина, после активации команды **01 → X**, никогда больше не перешла во внутреннее состояние 0 алгоритма $A$. Это требование ни в коей мере не влечет за собой каких-либо существенных ограничений на форму алгоритма[11]. (Нуль можно использовать только в командах-пустышках.)

[11] Более того, сам Тьюринг первоначально предполагал вообще *останавливать* машину всякий раз, когда она повторно переходит во внутреннее состояние «0» из любого другого состояния. В этом случае нам не только не понадобилось бы вышеупомянутое ограничение, мы спокойно могли бы обойтись и без команды **STOP**. Тем самым мы достигли бы существенного упрощения, поскольку последовательность **11110** в качестве команды нам была бы уже не нужна, и ее можно было бы использовать как разделитель, что позволило бы избавиться от последовательности **111110**. Это значительно сократило бы длину предписания $K$, и, кроме того, вместо пятеричной системы счисления мы обошлись бы четверичной.



Затем при определении алгоритма $A$ необходимо установить общее число $N$ внутренних состояний (включая и состояние 0, т. е. максимальное число внутренних состояний $A$ будет равно $N - 1$). Если в определении $A$ нет завершающей команды вида $(N-1)$**1 → Y**, то в конце следует добавить команду-пустышку $(N-1)$**1 → 0 0R**. Наконец, удалим из определения $A$ команду **01 → X** и добавим ее к приводимому ниже списку машинных команд, а каждый номер внутреннего состояния, фигурирующий в этом списке, увеличим на $N$ (символом ∅ обозначено результирующее внутреннее состояние 0, а символом «**X**» в записи «**11 → X**» представлена команда, которую мы рассмотрели выше). (В частности, первые две команды из списка примут в данном случае следующий вид: **01 →** $N$**1R**, $N$**0 →** $(N+4)$ **0R**.)

∅**1** → **01R**, **00** → **40R**, **01** → **01R**, **10** → **21R**,
**11** → **X**, **20** → **31R**, **21** → ∅**0R**, **30** → **551R**,
**31** → ∅**0R**, **40** → **40R**, **41** → **51R**, **50** → **40R**,
**51** → **61R**, **60** → **40R**, **61** → **71R**, **70** → **40R**,
**71** → **81R**, **80** → **40R**, **81** → **91R**, **90** → **100R**,
**91** → ∅**0R**, **100** → **111R**, **101** → ∅**0R**, **110** → **121R**,
**111** → **120R**, **120** → **131R**, **121** → **130R**, **130** → **141R**,
**131** → **140R**, **140** → **151R**, **141** → **10R**, **150** → **00R**,
**151** → ∅**0R**, **160** → **170L**, **161** → **161L**, **170** → **170L**,
**171** → **181L**, **180** → **170L**, **181** → **191L**, **190** → **170L**,
**191** → **201L**, **200** → **170L**, **201** → **211L**, **210** → **170L**,
**211** → **221L**, **220** → **220L**, **221** → **231L**, **230** → **220L**,
**231** → **241L**, **240** → **220L**, **241** → **251L**, **250** → **220L**,
**251** → **261L**, **260** → **220L**, **261** → **271L**, **270** → **321R**,
**271** → **281L**, **280** → **330R**, **281** → **291L**, **290** → **330R**,
**291** → **301L**, **300** → **330R**, **301** → **311L**, **310** → **330R**,
**311** → **110R**, **320** → **340L**, **321** → **321R**, **330** → **350R**,
**331** → **331R**, **340** → **360R**, **341** → **340R**, **350** → **371R**,
**351** → **350R**, **360** → **360R**, **361** → **381R**, **370** → **370R**,
**371** → **391R**, **380** → **360R**, **381** → **401R**, **390** → **370R**,

$$
\begin{array}{l}
391 \to 411\mathbf{R}, 400 \to 360\mathbf{R}, 401 \to 421\mathbf{R}, 410 \to 370\mathbf{R}, \\
411 \to 431\mathbf{R}, 420 \to 360\mathbf{R}, 421 \to 441\mathbf{R}, 430 \to 370\mathbf{R}, \\
431 \to 451\mathbf{R}, 440 \to 360\mathbf{R}, 441 \to 461\mathbf{R}, 450 \to 370\mathbf{R}, \\
451 \to 471\mathbf{R}, 460 \to 480\mathbf{R}, 461 \to 461\mathbf{R}, 470 \to 490\mathbf{R}, \\
471 \to 471\mathbf{R}, 480 \to 480\mathbf{R}, 481 \to 490\mathbf{R}, 490 \to 481\mathbf{R}, \\
491 \to 501\mathbf{R}, 500 \to 481\mathbf{R}, 501 \to 511\mathbf{R}, 510 \to 481\mathbf{R}, \\
511 \to 521\mathbf{R}, 520 \to 481\mathbf{R}, 521 \to 531\mathbf{R}, 530 \to 541\mathbf{R}, \\
531 \to 531\mathbf{R}, 540 \to 160\mathbf{L}, 541 \to \varnothing 0\mathbf{R}, 550 \to 531\mathbf{R}.
\end{array}
$$

Теперь мы готовы точно определить предельную длину предписания $K$, получаемого путем вышеприведенного построения, как функцию от длины алгоритма $A$. Сравним эту «длину» со «степенью сложности», определенной в § 2.6 (в конце комментария к возражению **Q8**). Для некоторой конкретной машины Тьюринга $T_m$ (например, той, что выполняет вычисление $A$) эта величина равна количеству знаков в двоичном представлении числа $m$. Для некоторого конкретного машинного *действия* $T_m(n)$ (например, выполнения предписания $K$) эта величина равна количеству двоичных цифр в большем из чисел $m$ и $n$. Обозначим через $\alpha$ и $\kappa$ количество двоичных цифр в $a$ и $k'$ соответственно, где

$$A = T_a \quad \text{и} \quad K = T_{k'}\,(= C_k).$$

Поскольку алгоритм $A$ содержит, как минимум, $2N - 1$ команд (учитывая, что первую команду мы исключили) и поскольку для каждой команды требуется, по крайней мере, три двоичные цифры, общее число двоичных цифр в номере его машины Тьюринга $a$ непременно должно удовлетворять условию

$$\alpha \geqq 6N - 6.$$

В вышеприведенном дополнительном списке команд для $K$ есть 105 мест (справа от стрелок), где к имеющемуся там числу следует прибавить $N$. Все получаемые при этом числа не превышают $N + 55$, а потому их расширенные двоичные представления содержат не более $2\log_2(N + 55)$ цифр, в результате чего общее количество двоичных цифр, необходимых для дополнительного определения внутренних состояний, не превышает $210\log_2(N + 55)$. Сюда нужно добавить цифры, необходимые для добавочных символов **0**, **1**, **R** и **L**, что составляет еще 527 цифр (включая одну возможную добавочную «команду-пустышку» и учитывая, что мы можем исключить шесть символов **0** по правилу, согласно которому **00** можно представить в виде **0**). Таким образом, для определения предписания $K$ требуется больше двоичных цифр, чем для определения алгоритма $A$, однако разница между этими двумя величинами не превышает $527 + 210\log_2(N + 55)$:

$$\kappa < \alpha + 527 + 210\log_2(N + 55).$$

Применив полученное выше соотношение $\alpha \geqq 6N - 6$, получим (учитывая, что $210\log_2 6 > 542$)

$$\kappa < \alpha - 15 + 210\log_2(\alpha + 336).$$

Затем найдем степень сложности $\eta$ конкретного вычисления $C_k(k)$, получаемого посредством этой процедуры. Вспомним, что степень сложности машины $T_m(n)$ определяется как количество двоичных цифр в большем из двух чисел $m$, $n$. В данной ситуации $C_k = T_k$, так что число двоичных цифр в числе «$m$» этого вычисления равно $\kappa$. Для того чтобы определить, сколько двоичных цифр содержит число «$n$» этого вычисления, рассмотрим ленту, содержащую вычисление $C_k(k)$. Эта лента начинается с последовательности символов **111110**, за которой следует двоичное выражение числа $k'$, и завершается последовательностью **11011111**. В соответствии с предложенным в НРК соглашением всю эту последовательность (без последней цифры) следует читать как двоичное число; эта операция дает нам номер «$n$», который присваивается ленте машины, выполняющей вычисление $T_m(n)$. То есть число двоичных цифр в данном конкретном номере «$n$» равно $\kappa + 13$, и, следовательно, число $\kappa + 13$ совпадает также со степенью сложности $\eta$ вычисления $C_k(k)$, благодаря чему мы можем записать $\eta = \kappa + 13 < \alpha - 2 + 210\log_2(\alpha + 336)$, или проще:

$$\eta < \alpha + 210\log_2(\alpha + 336).$$

Детали вышеприведенного рассуждения специфичны для данного конкретного предложенного еще в НРК способа кодирования машин Тьюринга, и при использовании какого-либо иного кодирования они также будут несколько иными. Основная же идея очень проста. Более того, прими мы формализм *λ-исчисления*, вся операция оказалась бы, в некотором смысле, почти

тривиальной. (Достаточно обстоятельное описание $\lambda$-исчисления Черча можно найти в НРК, конец главы 2; см. также [52].) Предположим, например, что алгоритм $A$ определяется некоторым $\lambda$-оператором **A**, выполняющим действие над другими операторами **P** и **Q**, что выражается в виде операции **(AP) Q**. Оператором **P** здесь представлено вычисление $C_p$, а оператором **Q** — число $q$. Далее, оператор **A** должен удовлетворять известному требованию, согласно которому для любых **P** и **Q** должно быть истинным следующее утверждение:

Если завершается операция **(AP) Q**, то операция **PQ** не завершается.

Мы без труда можем составить такую операцию $\lambda$-исчисления, которая не завершается, однако этот факт невозможно установить посредством оператора **A**. Например, положим

$$\mathbf{K} = \lambda\mathbf{x}.[(\mathbf{Ax})\,\mathbf{x}],$$

т. е. $\mathbf{KY} = (\mathbf{AY})\,\mathbf{Y}$ для любого оператора **Y**. Затем рассмотрим $\lambda$-операцию

**KK**.

Очевидно, что эта операция не завершается, поскольку $\mathbf{KK} = (\mathbf{AK})\,\mathbf{K}$, а завершение последней операции означало бы, что операция **KK** не завершается по причине принятой нами природы оператора **A**. Более того, оператор **A** не способен установить этот факт, потому что операция **(AK) K** не завершается. Если мы *полагаем*, что оператор **A** обладает требуемым свойством, то мы также должны *предположить*, что операция **KK** не завершается.

Отметим, что данная процедура дает значительную экономию. Если записать операцию **KK** в виде

$$\mathbf{KK} = \lambda\mathbf{y}.(\mathbf{yy})(\lambda\mathbf{x}.[(\mathbf{Ax})\,\mathbf{x}]),$$

то становится ясно, что число символов в записи операции **KK** всего на 16 больше аналогичного числа символов для алгоритма $A$ (если пренебречь точками, которые в любом случае избыточны)!

Строго говоря, это не совсем законно, поскольку в выражении для оператора **A** может также появиться и символ «**x**», и с этим нам придется что-то делать. Можно усмотреть сложность и в том, что генерируемое такой процедурой незавершающееся вычисление нельзя считать операцией над натуральными числами



(поскольку вторая **K** в записи **KK** «числом» не является). Вообще говоря, $\lambda$-исчисление не вполне подходит для работы с явными численными операциями, и зачастую бывает довольно сложно понять, каким образом ту или иную заданную алгоритмическую процедуру, применяемую к натуральным числам, можно выразить в виде операции $\lambda$-исчисления. По этим и подобным причинам обсуждение с привлечением машин Тьюринга имеет, как нам представляется, более непосредственное отношение к теме нашего исследования и достигает требуемого результата более наглядным путем.

# 3

# О НЕВЫЧИСЛИМОСТИ В МАТЕМАТИЧЕСКОМ МЫШЛЕНИИ

## 3.1. Гёдель и Тьюринг

В главе 2 была предпринята попытка продемонстрировать мощь и строгий характер аргументации в пользу утверждения (обозначенного буквой $\mathscr{G}$), суть которого заключается в том, что математическое понимание не может являться результатом применения какого-либо осмысленно осознаваемого и полностью достоверного алгоритма (или, что то же самое, алгоритмов; см. возражение **Q1**). В приводимых рассуждениях, однако, ни словом не упомянуто еще об одной возможности, существенно более серьезной и *ничуть не противоречащей* утверждению $\mathscr{G}$, а именно: убежденность математика в истинности своих выводов может оказаться результатом применения им некоего неизвестного и неосознаваемого алгоритма, или же, возможно, математик применяет какой-то вполне постижимый алгоритм, однако при этом не может знать наверняка (или хотя бы искренне верить), что выводы его являются целиком и полностью результатом применения этого самого алгоритма. Ниже я покажу, что, хотя подобные допущения и вполне приемлемы с логической точки зрения, вряд ли их можно счесть хоть сколько-нибудь правдоподобными.

Прежде всего следует указать на то, что тщательно выстраивая последовательности умозаключений (вполне, заметим, осознанных) с целью установления той или иной математической истины, математики вовсе не считают, что они лишь слепо следуют неким неосознаваемым правилам, будучи при этом не



в состоянии постичь эти правила ни рассудком, ни верой. Напротив, они твердо знают, что их аргументация опирается исключительно на непреложные истины — в основе своей существенно «очевидные»; столь же непреложными, на их взгляд, являются и все промежуточные умозаключения, составляющие упомянутую последовательность. Какой бы длинной, запутанной или даже концептуально неочевидной ни была цепь умозаключений, само рассуждение в основе своей остается принципиально неопровержимым и логически безупречным, а автор его искренне верит в свою правоту. Ни один математик не согласится с предположением о том, что на самом-то деле все его действия определяются какими-то совершенно иными процедурами, о которых он ничего не знает и в которые не верит, но которые, возможно, неким непостижимым образом исподволь влияют на его убеждения.

Разумеется, в этом отношении математики могут и ошибаться. Может быть, и впрямь существует какая-то алгоритмическая процедура, которая руководит всем математическим мышлением, оставаясь при этом неизвестной самим математикам. Всерьез принять такую возможность, пожалуй, легче людям, далеким от математики, нежели большинству из тех, для кого математика является профессией. Полагая, что деятельность математика не сводится к простому выполнению некоего неизвестного (и непостижимого) алгоритма (равно как и алгоритма, в существовании которого он испытывает сомнения), это самое большинство оказывается как нельзя более правым, в чем я и постараюсь убедить читателя в этой главе. Разумеется, полностью исключить возможность того, что суждения и убеждения математиков и в самом деле определяются какими-то неизвестными и неосознаваемыми факторами, нельзя; однако, даже если так оно и есть, я полагаю, что такие факторы не имеют ничего общего с алгоритмически описываемыми процедурами.

Весьма поучительным представляется рассмотреть точки зрения двух выдающихся мыслителей от математики, которым мы, собственно говоря, и обязаны идеями, приведшими нас к утверждению $\mathscr{G}$. Что, в самом деле, думал по этому поводу Гёдель? А Тьюринг? Примечательно, что, исходя из одинаковых математических данных, они пришли к противоположным, в сущности, выводам. Следует, впрочем, пояснить, что оба вывода находятся в полном согласии с утверждением $\mathscr{G}$. Гёдель, по

всей видимости, полагал, что разум, вообще говоря, не ограничен не только необходимостью выступать исключительно в качестве вычислительной сущности, но и конечными физическими параметрами самого мозга. Он даже упрекал Тьюринга за то, что тот не допускал такой возможности. По словам Хао Вана ([375], с. 326, см. также *Собрание сочинений* Гёделя, т. 2 [159], с. 297), соглашаясь с обоими, вытекающими из позиции Тьюринга положениями, т. е. с тем, что «мозг, в сущности, функционирует подобно цифровому компьютеру», и с тем, что «физические законы, равно как и наблюдаемые следствия из них, обладают конечным пределом точности», Гёдель напрочь отвергал утверждение Тьюринга о неотделимости разума от материи, считая это «свойственным эпохе предрассудком». Таким образом, согласно Гёделю, сам по себе *физический* мозг действует исключительно как вычислитель, разум же по отношению к мозгу представляет собой нечто высшее, вследствие чего активность разума оказывается свободной от ограничений, налагаемых вычислительными законами, управляющими поведением мозга как физического объекта. Гёдель, судя по его собственным словам(1), не считал, что утверждение $\mathscr{G}$ можно рассматривать в качестве *доказательства* его тезиса о невычислимости деятельности разума:

> «С другой стороны, учитывая доказанное ранее, следует допустить принципиальную возможность существования (и даже эмпирической реализации) некоей машины для доказательства теорем, каковая машина в сущности представляет собой эквивалент математической интуиции, однако *доказать* эту эквивалентность невозможно, как невозможно доказать и то, что на выходе такой машины мы будем получать только *корректные* теоремы конечной теории чисел».

Надо сказать, что вышеприведенное допущение ни в коей мере не противоречит $\mathscr{G}$ (и я ничуть не сомневаюсь, что Гёделю был хорошо известен тот недвусмысленный вывод, какой в моей формулировке получил обозначение $\mathscr{G}$). Гёдель допускал *логическую возможность* того, что разум математика может функционировать в соответствии с некоторым алгоритмом, о котором сам математик не знает, либо знает, но в таком случае не может быть однозначно уверен в его обоснованности (... *доказать* ... невозможно, ... только *корректные* теоремы ...). В соответствии с моей собственной терминологией такой алгоритм следует отнести к категории «непознаваемо обоснованных». Разумеется, совсем иное дело действительно *поверить* в возможность того, что деятельность разума математика и в самом деле определяется таким вот непознаваемо обоснованным алгоритмом. Похоже, сам Гёдель в это так и не поверил — и оказался в результате окружен компанией мистиков (точка зрения $\mathscr{D}$), которые полагают, что средствами науки о феноменах физического мира разум объяснить невозможно.



Что же касается Тьюринга, то он, по-видимому, мистическую точку зрения не принял, будучи в то же время солидарен с Гёделем в том, что мозг, как и всякий другой физический объект, должен функционировать каким-либо вычислимым образом (вспомним о «тезисе Тьюринга», § 1.6). Таким образом, Тьюрингу пришлось искать какой-то другой способ обойти затруднение в виде утверждения $\mathscr{G}$. При этом особенно значимым ему показался тот факт, что математикам-людям свойственно делать ошибки; если мы хотим, чтобы наш компьютер стал подлинно разумным, следует позволить ему хоть иногда ошибаться(2):

> «Иными словами, это означает, что если мы требуем от машины непогрешимости, то не стоит ожидать от нее еще и разумности. Существует несколько теорем, суть которых почти буквально сводится к вышеприведенному утверждению. Однако в этих теоремах ничего не говорится о степени разумности, которую нам может продемонстрировать машина, не претендующая на непогрешимость».

Под «теоремами» Тьюринг, вне всякого сомнения, подразумевает теорему Гёделя и другие аналогичные теоремы — такие, например, как его собственная, «вычислительная» версия теоремы Гёделя. То есть, по Тьюрингу, получается, что наиболее существенной способностью человеческого математического мышления является способность ошибаться, благодаря которой свойственное (предположительно) разуму неточно-алгоритмическое функционирование обеспечивает бóльшую мощность, нежели возможно получить посредством каких угодно полностью обоснованных алгоритмических процедур. Исходя из этого допущения, Тьюринг предложил способ обойти ограничение, налагаемое следствиями из теоремы Гёделя: мыслительная де-

ятельность математика подчиняется-таки некоему алгоритму, только не «непознаваемо обоснованному», а формально необоснованному. Таким образом, точка зрения Тьюринга приходит в полное согласие с утверждением $\mathscr{G}$, а сам Тьюринг, по-видимому, присоединяется к сторонникам точки зрения $\mathscr{A}$.

Завершая дискуссию, я хотел бы представить мои собственные причины усомниться в том, что «необоснованность» управляющего разумом математика алгоритма может послужить *подлинным* объяснением тому, что в этом самом разуме происходит. Как бы ни обстояло дело в действительности, в самой идее о том, что превосходство человеческого разума над точной машиной достигается за счет *неточности* разума, мне видится какое-то глубинное противоречие, особенно когда речь — как в нашем случае — идет о способности математика *открывать неопровержимые математические истины*, а не о его оригинальности или творческих способностях. Поразительно, что два великих мыслителя, какими, несомненно, являются Гёдель и Тьюринг, руководствуясь соображениями вроде утверждения $\mathscr{G}$, пришли к выводам (пусть и различным), которые многие из нас склонны считать, скажем так, маловероятными. Кроме того, весьма интересно поразмыслить о том, к каким бы выводам они пришли, имей они шанс хоть сколько-нибудь всерьез предположить, что физический процесс может иногда оказаться в основе своей невычислимым — в соответствии с точкой зрения $\mathscr{C}$, ради продвижения которой и была написана эта книга.

В последующих разделах (особенно, в §§ 3.2–3.22) я представлю вашему вниманию несколько детальных обоснований (некоторые из них довольно сложны, запутаны или специальны), целью которых является демонстрация неспособности вычислительных моделей $\mathscr{A}$ и $\mathscr{B}$ выступить в качестве вероятной основы для исследования феномена математического понимания. Если читатель не нуждается в подобном убеждении либо не склонен погружаться в детали, то я бы порекомендовал ему (или ей) все же начать чтение, а затем, когда уж совсем надоест, переходить сразу к итоговому воображаемому диалогу (§ 3.23). Если у вас затем появится желание вернуться к пропущенным рассуждениям, буду только рад, если же нет — забудьте о них и читайте дальше.



## 3.2. Способен ли необоснованный алгоритм познаваемым образом моделировать математическое понимание?

Согласно выводу $\mathscr{G}$, для того чтобы математическое понимание могло оказаться результатом выполнения некоего алгоритма, этот алгоритм должен быть необоснованным или непознаваемым, если же он сам по себе обоснован и познаваем, то о его обоснованности должно быть принципиально невозможно узнать наверняка (такой алгоритм мы называем непознаваемо обоснованным); кроме того, возможно, что различные математики «работают» на различных типах таких алгоритмов. Под «алгоритмом» здесь понимается просто какая-нибудь вычислительная процедура (см. § 1.5), т. е. любой набор операций, который можно, в принципе, смоделировать на универсальном компьютере с неограниченным объемом памяти. (Как нам известно из обсуждения возражения **Q8**, § 2.6, «неограниченность» объема памяти в данном идеализированном случае на результаты рассуждения никак не влияет.) Такое понятие алгоритма включает в себя нисходящие процедуры, восходящие самообучающиеся системы, а также различные их сочетания. Сюда, например, входят любые процедуры, которые можно реализовать с помощью искусственных нейронных сетей (см. § 1.5). Этому определению отвечают и иные типы восходящих механизмов — например, так называемые «генетические алгоритмы», повышающие свою эффективность с помощью некоей встроенной процедуры, аналогичной дарвиновской эволюции (см. § 3.11).

О специфике приложения аргументации, представляемой в настоящем разделе (равно как и доводов, выдвинутых в главе 2), к восходящим процедурам я еще буду говорить в §§ 3.9–3.22 (краткое изложение их можно найти в воображаемом диалоге, § 3.23). Пока же, для большей ясности изложения, будем рассуждать, исходя из допущения, что в процессе участвует один-единственный тип алгоритмических процедур, а именно — нисходящие. Такую алгоритмическую процедуру можно относить как к отдельному математику, так и к математическому сообществу в целом. В комментариях к возражениям **Q11** и **Q12**, § 2.10, рассматривалось предположение о том, что разным людям могут быть свойственны *различные* обоснованные и известные алгоритмы, причем мы пришли к заключению, что такая возможность

не влияет на результаты рассуждения сколько-нибудь значительным образом. Возможно также, что разные люди постигают истину посредством различных *необоснованных* и *непознаваемых* алгоритмов; к этому вопросу мы вернемся несколько позже (см. §3.7). А пока, повторюсь, будем считать, что в основе математического понимания лежит одна-единственная алгоритмическая процедура. Можно, кроме того, ограничить рассматриваемую область той частью математического понимания, которая отвечает за доказательство $\Pi_1$-высказываний (т. е. определений тех операций машины Тьюринга, которые не завершаются; см. комментарий к возражению **Q10**). В дальнейшем вполне достаточно интерпретировать сочетание «математическое понимание» как раз в таком, ограниченном смысле (см. формулировку $\mathscr{G}^{**}$, с. 166).

В зависимости от познаваемости предположительно лежащей в основе математического понимания алгоритмической процедуры $F$ (будь то обоснованной или нет), следует четко выделять три совершенно различных случая. Процедура $F$ может быть:

**I** сознательно познаваемой, причем познаваем также и тот факт, что именно эта алгоритмическая процедура ответственна за математическое понимание;

**II** сознательно познаваемой, однако тот факт, что математическое понимание основывается именно на этой алгоритмической процедуре, остается как неосознаваемым, так и непознаваемым;

**III** неосознаваемой и непознаваемой.

Рассмотрим сначала полностью сознательный случай **I**. Поскольку и сам алгоритм, и его роль являются познаваемыми, мы вполне можем счесть, что мы о них *уже* знаем. В самом деле, ничто не мешает нам вообразить, что все наши рассуждения имеют место уже после того, как мы получили в наше распоряжение соответствующее знание — ведь слово «познаваемый» как раз и подразумевает, что такое время, по крайней мере, в принципе, когда-нибудь да наступит. Итак, алгоритм $F$ нам известен, при этом известна и его основополагающая роль в математическом понимании. Как мы уже видели (§2.9), такой алгоритм эффективно эквивалентен формальной системе $\mathbb{F}$. Иными словами, получается, что математическое понимание — или хотя бы понимание



математики каким-то отдельным математиком — эквивалентно выводимости в рамках некоторой формальной системы $\mathbb{F}$. Если мы хотим сохранить хоть какую-то надежду удовлетворить выводу $\mathscr{G}$, к которому нас столь неожиданно привели изложенные в предыдущей главе соображения, то придется предположить, что система $\mathbb{F}$ является *необоснованной*. Однако, как это ни странно, необоснованность в данном случае ситуацию ничуть не меняет, поскольку, в соответствии с **I**, известная формальная система $\mathbb{F}$ является действительно *известной*, то есть любой математик *знает* и, как следствие, *верит*, что именно эта система лежит в основе его (или ее) математического понимания. А такая вера автоматически влечет за собой веру (пусть и ошибочную) в обоснованность системы $\mathbb{F}$. (Согласитесь, крайне неразумно выглядит точка зрения, в соответствии с которой математик позволяет себе не верить в самые фундаментальные положения собственной заведомо неопровержимой системы взглядов.) Независимо от того, является ли система $\mathbb{F}$ действительно обоснованной, *вера* в ее обоснованность уже содержит в себе веру в то, что утверждение $G(\mathbb{F})$ (или, как вариант, $\Omega(\mathbb{F})$, см. §2.8) истинно. Однако, поскольку теперь мы полагаем (исходя из веры в справедливость теоремы Гёделя), что истинность утверждения $G(\mathbb{F})$ в рамках системы $\mathbb{F}$ недоказуема, это противоречит предположению о том, что система $\mathbb{F}$ является основой *всякого* (существенного для рассматриваемого случая) математического понимания. (Это соображение одинаково справедливо как для отдельных математиков, так и для всего математического сообщества в целом; его можно применять индивидуально к любому из всевозможных алгоритмов, предположительно составляющих основу мыслительных процессов того или иного математика. Более того, согласно предварительной договоренности, для нас на данный момент важна применимость этого соображения лишь в той области математического понимания, которая имеет отношение к доказательству $\Pi_1$-высказываний.) Итак, невозможно знать наверняка, что некий гипотетический известный необоснованный алгоритм $F$, предположительно лежащий в основе математического понимания, и в самом деле выполняет эту роль. Следовательно, случай **I** исключается, независимо от того, является система $\mathbb{F}$ обоснованной или нет. Если система $\mathbb{F}$ сама по себе познаваема, то следует рассмотреть возможность **II**, суть которой заключается в том, что система $\mathbb{F}$ все же может составлять основу

математического понимания, однако узнать об этой ее роли мы не в состоянии. Остается в силе и возможность **III**: сама система $\mathbb{F}$ является как неосознаваемой, так и непознаваемой.

На данный момент мы достигли следующего результата: случай **I** (по крайней мере, в контексте полностью нисходящих алгоритмов) как сколько-нибудь серьезную возможность рассматривать нельзя; тот факт, что система $\mathbb{F}$ может в действительности оказаться и необоснованной, как выяснилось, сути проблемы ничуть не меняет. Решающим фактором здесь является невозможность точно установить, является та или иная гипотетическая система $\mathbb{F}$ (независимо от ее обоснованности) основой для формирования математических убеждений или же нет. Дело не в непознаваемости самого алгоритма, но в непознаваемости того факта, что процесс понимания *действительно* происходит в соответствии с данным алгоритмом.

## 3.3. Способен ли познаваемый алгоритм непознаваемым образом моделировать математическое понимание?

Перейдем к случаю **II** и попытаемся серьезно рассмотреть возможность того, что математическое понимание на деле эквивалентно некоторому сознательно познаваемому алгоритму либо формальной системе, однако эквивалентность эта принципиально непознаваема. Иными словами, даже при условии познаваемости той или иной гипотетической формальной системы $\mathbb{F}$ мы никоим образом не можем убедиться в том, что именно эта конкретная система действительно лежит в основе нашего математического понимания. Правдоподобно ли такое предположение?

Если упомянутая гипотетическая формальная система $\mathbb{F}$ не является *уже* известной, то в этом случае нам, как и ранее, следует полагать, что она может, по крайней мере, в принципе, когда-нибудь таковой стать. Вообразим, что этот светлый день наконец наступил, и допустим, что в нашем распоряжении имеется точное и подробное описание этой самой системы. Предполагается, что формальная система $\mathbb{F}$, будучи, возможно, крайне замысловатой, все же достаточно проста для того, чтобы мы оказались способны, по крайней мере, в принципе, постичь ее на вполне сознательном уровне. При этом нам не позволено испытывать *уверенность* в том, что система $\mathbb{F}$ действительно целиком и полностью



охватывает всю совокупность наших твердых математических убеждений и интуитивных озарений (по крайней мере в том, что касается $\Pi_1$-высказываний). Это (вообще-то вполне логичное) предположение оказывается на деле в высшей степени неправдоподобным, в причинах чего мы и попытаемся разобраться. Более того, несколько позднее я покажу, что даже будь оно истинным, это не принесло бы никакой радости тем ИИ-энтузиастам, которые видят смысл жизни в создании робота-математика. Мы еще поговорим об этом в конце данного раздела и — более подробно — в §§ 3.15 и 3.29.

Дабы подчеркнуть тот факт, что существование подобной системы $\mathbb{F}$ и в самом деле следует полагать *логически* возможным, вспомним о «машине для доказательства теорем», возможности создания которой, согласно Гёделю, логически исключить нельзя (см. цитату в § 3.1). В сущности, такую «машину», как я поясню ниже, как раз и можно представить в виде некоторой алгоритмической процедуры $F$, соответствующей вышеприведенным пунктам **II** или **III**. Как отмечает Гёдель, его гипотетическая машина для доказательства теорем может быть «эмпирически реализована», что соответствует требованию «сознательной познаваемости» процедуры $F$ в случае **II**; если же подобная реализация оказывается невозможной, то мы, по сути, имеем дело со случаем **III**.

На основании своей знаменитой теоремы Гёдель утверждал, что невозможно *доказать* «эквивалентность» процедуры $F$ (или, что то же самое, формальной системы $\mathbb{F}$; см. § 2.9) «математической интуиции» (см. ту же цитату). В определении случая **II** (и, как следствие, **III**) я сформулировал это фундаментальное ограничение, налагаемое на $\mathbb{F}$, несколько по-иному: «Тот факт, что математическое понимание основывается именно на этой алгоритмической процедуре, остается как неосознаваемым, так и непознаваемым».

Это ограничение (необходимость в котором следует из обоснованного в § 3.2 исключения случая **I**) со всей очевидностью приводит к невозможности показать, что процедура $F$ эквивалентна математической интуиции, поскольку посредством подобной демонстрации мы могли бы однозначно убедиться в том, что процедура $F$ действительно выполняет ту роль, о самом факте выполнения которой мы предположительно не в состоянии ничего знать. И наоборот, если бы эта самая роль процедуры $F$ (роль фундаментального алгоритма, в соответствии с которым

осуществляется постижение математических истин) допускала осознанное познание (в том смысле, что мы могли бы в полной мере постичь, как именно процедура $F$ выполняет эту свою роль), то нам пришлось бы признать и обоснованность $F$. Ибо если мы не допускаем, что процедура $F$ целиком и полностью обоснованна, то это означает, что мы отвергаем какие-то ее следствия. А ее следствиями являются как раз те математические положения (или хотя бы только $\Pi_1$-высказывания), которые мы полагаем-таки истинными. Таким образом знание роли процедуры $F$ равнозначно наличию *доказательства* $F$, хотя такое «доказательство» и нельзя считать формальным доказательством в рамках некоторой заранее заданной формальной системы.

Отметим также, что истинные $\Pi_1$-высказывания можно рассматривать в качестве примеров тех самых «корректных теорем конечной теории чисел», о которых говорил Гёдель. Более того, если понятие «конечной теории чисел» включает в себя $\mu$-операцию «отыскания наименьшего натурального числа, обладающего таким-то свойством», в каковом случае оно включает в себя и процедуры, выполняемые машинами Тьюринга (см. конец § 2.8), то тогда частью конечной теории чисел следует считать все $\Pi_1$-высказывания. Иными словами, получается, что доказательство гёделевского типа не дает четкого способа исключить из рассмотрения случай **II**, руководствуясь одними лишь строго логическими основаниями — по крайней мере, до тех пор, пока мы полагаем, что Гёдель был прав.

С другой стороны, можно задаться вопросом об общем *правдоподобии* предположения **II**. Рассмотрим, что повлечет за собой существование познаваемой процедуры $F$, непознаваемым образом эквивалентной человеческому математическому пониманию (заведомо непогрешимому). Как уже отмечалось, ничто не мешает нам мысленно перенестись в некое будущее время, в котором эта процедура окажется обнаружена и подробно описана. Известно также (см. § 2.7), что формальная система задается в виде некоторого набора *аксиом* и *правил действия*. *Теоремы* системы $\mathbb{F}$ представляют собой утверждения (иначе называемые «положениями»), выводимые из аксиом с помощью правил действия, причем все теоремы можно сформулировать посредством того же набора символов, который используется для выражения аксиом. А теперь представим себе, что теоремы системы $\mathbb{F}$ в точности совпадают с теми положениями (сформулированными с помощью упомянутых символов), неопровержимую истинность которых математики, *в принципе*, способны самостоятельно установить.



Допустим на минуту, что перечень аксиом системы $\mathbb{F}$ является *конечным*. Сами же аксиомы суть не что иное, как частные случаи соответствующих теорем. Однако неопровержимую истинность каждой теоремы мы можем, в принципе, постичь посредством математического понимания и интуиции. Следовательно, каждая аксиома в отдельности должна выражать нечто такое, что (по крайней мере, в принципе) постижимо посредством этого самого математического понимания. Иными словами, для каждой отдельной аксиомы когда-нибудь непременно настанет (либо *принципиально* возможно, что настанет) время, когда ее неопровержимая истинность будет однозначно установлена. Так, рассматривая одну за другой, мы сможем устанавливать истинность любой отдельно взятой аксиомы системы $\mathbb{F}$. Таким образом, в конечном итоге будет установлена (либо *принципиально* возможно, что будет установлена) неопровержимая истинность всех отдельно взятых аксиом. Соответственно, настанет время, когда будет установлена неопровержимая истинность всей совокупности аксиом системы $\mathbb{F}$ в целом.

А как быть с правилами действия? Можем ли мы предположить, что настанет время, когда будет однозначно установлена неопровержимая обоснованность этих правил? Во многих формальных системах правилами действия служат достаточно простые утверждения, каждое из которых с очевидностью «неопровержимо», например: «Если установлено, что высказывание $P$ является теоремой и высказывание $P \Rightarrow Q$ является теоремой, то можно заключить, что высказывание $Q$ также является теоремой» (относительно символа $\Rightarrow$ «следует» см. НРК, с. 393, или [223]). Признать неоспоримую справедливость таких правил совсем не трудно. С другой стороны, среди правил действия встречаются и гораздо более тонкие отношения, справедливость которых вовсе не так очевидна; прежде чем прийти к однозначному решению относительно того, считать то или иное такое правило «неопровержимо обоснованным» или нет, нам, возможно, потребуется прибегнуть к весьма подробному и тщательному анализу. Более того, как мы вскоре убедимся, в наборе правил действия формальной системы $\mathbb{F}$ неизбежно имеются такие правила, неоспоримая обоснованность которых не может быть достоверно

установлена ни одним математиком — причем мы все еще полагаем, что число аксиом в системе $\mathbb{F}$ конечно.

В чем же причина? Перенесемся в воображении в то самое время, когда уже однозначно установлена неопровержимая справедливость всех аксиом формальной системы $\mathbb{F}$. Перед нами открывается замечательная возможность без помех рассмотреть всю систему $\mathbb{F}$ целиком. Попробуем допустить, что все правила действия системы $\mathbb{F}$ можно также считать справедливыми безо всяких оговорок. Хотя предполагается, что мы еще не можем знать наверняка, что система $\mathbb{F}$ действительно включает в себя всю математику, которая в принципе доступна человеческому пониманию и интуиции, мы должны к настоящему моменту уже убедиться в том, что система $\mathbb{F}$ является, по меньшей мере, неоспоримо обоснованной, поскольку справедливость как ее аксиом, так и ее правил действия безоговорочно нами принимается. Следовательно, мы также должны уже быть уверены в том, что система $\mathbb{F}$ *непротиворечива*. Не забываем, разумеется, и о том, что, в силу этой непротиворечивости, утверждение $G(\mathbb{F})$ также должно быть истинным — более того, *неопровержимо* истинным! Однако, поскольку предполагается, что система $\mathbb{F}$ фактически (хотя нам об этом неизвестно) включает в себя всю совокупность того, что безоговорочно доступно нашему пониманию, утверждение $G(\mathbb{F})$ должно на деле представлять собой теорему системы $\mathbb{F}$. Согласно теореме Гёделя, такое, вообще говоря, возможно только в том случае, если формальная система $\mathbb{F}$ *противоречива*. Если же система $\mathbb{F}$ противоречива, то одной из теорем этой системы является утверждение «$1 = 2$». Следовательно, утверждение «$1 = 2$» должно быть, в принципе, доступно нашему математическому пониманию — очевидное противоречие!

Несмотря на это, следует, по крайней мере, учесть саму *возможность* того, что математики действуют (не зная о том) в рамках системы $\mathbb{F}$, которая является, по существу, *необоснованной*. К этому вопросу я еще вернусь в § 3.4, пока же (в пределах данного раздела) будем полагать, что на самом деле процедуры, лежащие в основе математического понимания, целиком и полностью обоснованны. При данных обстоятельствах, если мы продолжаем настаивать на том, что все правила действия нашей формальной системы $\mathbb{F}$ с конечным набором аксиом безоговорочно истинны, нам остается лишь признать, что противоречие действительно имеет место. Следовательно, среди правил действия системы $\mathbb{F}$



должно быть по крайней мере одно правило, обоснованность которого не может неопровержимо установить ни один математик (хотя в действительности это правило является обоснованным).

Все вышеприведенные рассуждения опирались на то допущение, что система $\mathbb{F}$ задается конечным набором аксиом. В качестве возможного альтернативного решения можно предположить, что количество аксиом в системе $\mathbb{F}$ бесконечно. Относительно этой возможности необходимо сделать некоторые комментарии. Для того чтобы систему $\mathbb{F}$ можно было определить как формальную в требуемом смысле — т. е. как систему, в рамках которой всегда можно однозначно установить (посредством некоторой заранее заданной вычислительной процедуры), что предполагаемое доказательство того или иного положения действительно является доказательством в соответствии с правилами системы, — необходимо, чтобы ее бесконечный набор аксиом можно было выразить каким-то конечно определяемым образом. Вообще говоря, всегда допускается некоторая свобода в отношении выбора конкретного способа представления формальной системы, в соответствии с которым операции системы определяются либо как аксиомы, либо как правила действия. Так, стандартная аксиоматическая система теории множеств — система Цермело—Френкеля (обозначаемая здесь как $\mathbb{ZF}$) — включает в себя бесконечное количество аксиом, выражаемых посредством структур, называемых «схемами аксиом». Путем соответствующего переформулирования систему $\mathbb{ZF}$ можно выразить таким образом, что количество действительных аксиом станет конечным[3]. Более того, действуя определенным образом, такое можно проделать с любой схемой аксиом, являющейся «формальной» в требуемом нами вычислительном смысле[1].

Может создаться впечатление, что вышеприведенное рассуждение (целью которого является исключение из списка возможных вариантов случая **II**) применимо к любой (обоснованной) системе $\mathbb{F}$, вне зависимости от того, конечно или бесконечно количество ее аксиом. Это и в самом деле так, однако в процессе приведения бесконечной схемы аксиом к конечному виду мы можем ввести новые правила действия, которые могут оказаться не

[1] Одним из достаточно тривиальных «подходов», с помощью которых можно осуществить упомянутое переформулирование, является следующий: нужно просто принять за набор правил действия требуемой системы последовательность операций машины Тьюринга, корректно реализующей алгоритм $F$.

столь самоочевидно обоснованными. Так, представляя себе, в соответствии с вышеизложенными соображениями, времена, когда нам станут известны все аксиомы и правила действия системы $\mathbb{F}$ (при этом также предполагается, что все теоремы этой гипотетической системы в точности совпадают с теоремами, которые в принципе доступны человеческим пониманию и интуиции), мы никоим образом не можем быть уверены в принципиальной возможности неопровержимого установления обоснованности правил действия такой системы $\mathbb{F}$, в отличие от ее аксиом (даже если эти правила действительно являются обоснованными). Дело в том, что, в отличие от аксиом, правила действия не принадлежат к теоремам формальной системы. Мы же полагаем, что неопровержимо установить можно лишь обоснованность *теорем* системы $\mathbb{F}$.

Не совсем ясно, возможно ли продолжить данное рассуждение, оставаясь при этом в рамках строгой логики. Если мы полагаем справедливой возможность **II**, то нам приходится признать, что существует некая формальная система $\mathbb{F}$ (на основании которой человек постигает истинность $\Pi_1$-высказываний), целиком и полностью понимаемая математиками, обладающая конечным набором аксиом, справедливость которых не вызывает никаких сомнений, и конечной системой правил действия $\mathscr{R}$, которая, впрочем, содержит по крайней мере одну операцию, полагаемую фундаментально сомнительной. Каждая отдельно взятая теорема системы $\mathbb{F}$ неизбежно оказывается утверждением, истинность которого может быть неопровержимо установлена, — что, собственно говоря, удивительно, учитывая тот факт, что многие из этих теорем выводятся с помощью сомнительных правил системы $\mathscr{R}$. Кроме того, хотя математик и может (в принципе) установить истинность каждой из упомянутых теорем *в отдельности, единообразной* процедуры для этого не существует. Можно ограничить область рассмотрения теми теоремами системы $\mathbb{F}$, которые представляют собой $\Pi_1$-высказывания. Применяя сомнительную систему правил $\mathscr{R}$, мы можем вычислительным способом сгенерировать перечень тех $\Pi_1$-высказываний, справедливость которых может быть однозначно установлена математиками. В конечном счете, человек, воспользовавшись пониманием и интуицией, оказывается способен установить справедливость каждого из этих $\Pi_1$-высказываний в отдельности. Однако в каждом конкретном случае для такого установления применяются



методы рассуждений, существенно отличающиеся от правила $\mathscr{R}$, с помощью которого было получено данное $\Pi_1$-высказывание. Раз за разом нам приходится добавлять в систему все новые, все более изощренные плоды человеческого разума — с тем, чтобы можно было неопровержимо доказать истинность каждого последующего $\Pi_1$-высказывания. Словно по волшебству, истинными оказываются все $\Pi_1$-высказывания, впрочем истинность некоторых из них можно установить лишь после привлечения какого-либо фундаментально нового метода рассуждения, причем необходимость в этом возникает вновь и вновь, на все более глубоких уровнях. Более того, *любое* $\Pi_1$-высказывание, неоспоримую истинность которого можно установить — причем неважно, каким методом, — оказывается уже включенным в тот самый перечень, который мы сгенерировали ранее с помощью системы правил $\mathscr{R}$. Наконец, существует еще и особое *истинное* $\Pi_1$-высказывание $G(\mathbb{F})$, которое явным образом выводится из знания формальной системы $\mathbb{F}$, однако истинность которого *не может* быть неопровержимо установлена ни одним математиком. В лучшем случае, математик сможет понять, что истинность $G(\mathbb{F})$ непосредственно обусловлена обоснованностью сомнительной системы правил действия $\mathscr{R}$, которая, по всей видимости, обладает некоей чудесной способностью определять, истинность каких именно $\Pi_1$-высказываний *может* быть неопровержимо установлена человеком.

Могу себе представить, что кому-то все это, возможно, покажется не *совсем* бессмысленным. Ко многим своим выводам математики приходят на основании предпосылок, которые можно назвать «эвристическими принципами» — такой принцип не дает непосредственного *доказательства* предполагаемого вывода, однако дает основания ожидать, что истинным неизбежно окажется именно такой вывод. Собственно доказательство может быть получено и позднее, причем совершенно иными методами. Мне, однако, представляется, что подобные эвристические принципы имеют на деле очень мало общего с нашей гипотетической системой правил $\mathscr{R}$. В сущности, такие принципы способны лишь углубить наше сознательное понимание причин, в соответствии с которыми оказывается истинным тот или иной математический вывод[2]. Впоследствии, в результате более серьезной разработ-

[2] Эвристический принцип такого рода может принять форму гипотезы — в качестве примера укажем весьма значительную гипотезу Таниямы (обобщенную

ки соответствующих математических методов, часто становится вполне ясно, почему именно сработал тот или иной эвристический принцип. В большинстве же случаев вполне проясняется лишь один вопрос: при каких именно *обстоятельствах* данный эвристический принцип гарантированно работает, а при каких — нет; иначе говоря, если не соблюдать известной осторожности, можно прийти к весьма и весьма ошибочным выводам. Если же осторожность соблюдена, сам такой принцип становится чрезвычайно мощным и надежным инструментом математического доказательства. Он не снабдит вас сверхъестественно достоверной алгоритмической процедурой для установления справедливости $\Pi_1$-высказываний, причины успешного функционирования которой будут принципиально недоступны человеческому пониманию; вместо этого он предоставит средства для углубления вашего математического понимания и усиления вашей же интуиции. А в этом, согласитесь, есть нечто, в корне отличное от алгоритма $F$ (или формальной системы $\mathbb{F}$), описанного в соответствии с возможностью **II**. Более того, никто никогда и не предлагал эвристического принципа, позволившего бы сгенерировать в точности *все* $\Pi_1$-высказывания, истинность которых может быть однозначно установлена математиками.

Разумеется, из всего этого вовсе не следует, что упомянутый алгоритм $F$ (гипотетическая машина Гёделя для доказательства теорем) является логически невозможным; однако, с позиции нашего математического понимания, вероятность существования такой машины представляется исключительно малой. Во всяком случае, в настоящее время ни у кого пока нет ни малейшего предположения относительно возможной природы подобного алгоритма $F$, равно как нет и никаких намеков на его действительное существование. Он может существовать, в лучшем случае, в качестве *гипотезы* — причем гипотезы недоказуемой. (Ее доказательство будет равносильно ее опровержению!) Мне думается, что со стороны любого из сторонников идеи ИИ (независимо от

позднее в так называемую «философскую теорию Лэнгленда»), в виде следствия из которой можно представить самое, пожалуй, знаменитое из $\Pi_1$-высказываний, известное широкой публике как «последняя теорема Ферма» (см. также примечание к с. 318). Однако рассуждение, предложенное Эндрю Уайлзом в качестве доказательства утверждения Ферма, представляет собой не рассуждение, независимое от гипотезы Таиямы, — каким оно неизбежно оказалось бы, будь эта гипотеза правилом системы «$\mathscr{R}$», — но рассуждение, *доказывающее* (в соответствующем случае) саму гипотезу Таиямы!



того, принадлежит он к лагерю $\mathscr{A}$ или $\mathscr{B}$) является в высшей степени безрассудным возлагать какие бы то ни было надежды на отыскание такой алгоритмической процедуры[3] (обобщенной здесь в виде алгоритма $F$), само существование которой крайне сомнительно, а точное построение (существуй она в действительности) едва ли по силам любому из ныне живущих математиков или логиков.

Можно ли допустить, что подобный алгоритм $F$ все же существует и, более того, может быть получен с помощью достаточно сложных вычислительных процедур восходящего типа? В §§ 3.5—3.23, в рамках обсуждения случая **III**, я приведу серьезные логические доводы, убедительно демонстрирующие, что ни одна из познаваемых восходящих процедур не в состоянии привести нас к алгоритму $F$, даже если бы он и в самом деле существовал. Таким образом, можно заключить, что в качестве сколько-нибудь серьезной логической возможности нельзя рассматривать даже «гёделеву машину для доказательства теорем» — если, конечно, не допустить, что в основе всего математического понимания в целом лежат некие «непознаваемые механизмы», природа которых, увы, не оставляет поборникам ИИ ни единого шанса.

Прежде чем мы перейдем к обещанному более подробному обсуждению случая **III**, необходимо разобраться до конца со случаем **II** — здесь остается еще одна альтернатива, суть которой заключается в том, что фундаментальная алгоритмическая процедура $F$ (или формальная система $\mathbb{F}$) может оказаться *необоснованной* (случай **I**, как мы помним, такой лазейки не допускал). Может ли быть так, что человеческое математическое понимание представляет собой эквивалент некоего познаваемого алгоритма, который в основе своей ошибочен? Рассмотрим эту возможность подробнее.

[3] Мне, разумеется, могут возразить, и не без оснований, что создание робота-математика отнюдь не входит в перечень ближайших задач исследований в области искусственного интеллекта; соответственно, попытки отыскания упомянутого алгоритма $F$ следует полагать преждевременными либо вовсе ненужными. Такое возражение, однако, может означать лишь то, что возражающий не совсем ясно представляет себе цели и суть настоящего обсуждения. Те точки зрения, согласно которым человеческий интеллект в целом объясним посредством алгоритмических процессов, неявно подразумевают, что алгоритм $F$ — познаваемый или нет — потенциально существует; к нашему же выводу мы пришли, всего лишь применив свой интеллект. Математические способности не являются в этом отношении чем-то особенным; см., в частности, §§ 1.18, 1.19.

## 3.4. Не действуют ли математики, сами того не осознавая, в соответствии с необоснованным алгоритмом?

Допустим, что в основе математического понимания и в самом деле лежит некая необоснованная формальная система $\mathbb{F}$. Как же мы тогда можем быть уверены, что наши математические представления в отношении того, что считать неоспоримо истинным, не введут нас в один прекрасный день в какое-нибудь фундаментальное заблуждение? А может, это уже случилось? Ситуация несколько отличается от той, что рассматривалась в связи со случаем **I**, где мы исключили возможность нашего *знания* о том, что некая система $\mathbb{F}$ и в самом деле является необоснованной. Здесь же мы допускаем, что подобная *роль* системы $\mathbb{F}$ принципиально непознаваема, вследствие чего нам придется повторно рассмотреть вариант с возможной необоснованностью $\mathbb{F}$. Можно ли считать действительно правдоподобным предположение о том, что фундаментом для наших неопровержимых математических убеждений служит некая необоснованная система — настолько необоснованная, что одним из этих убеждений может, в принципе, оказаться уверенность в истинности равенства $1 = 2$. Несомненно одно: если мы не можем доверять собственным *математическим* суждениям, то мы равным образом не можем доверять *и всем остальным* своим суждениям об устройстве и функционировании окружающего нас мира, поскольку математические суждения составляют весьма существенную часть всего нашего научного понимания.

Кто-то, тем не менее, возразит, что нет ничего *невероятного* в том, что какие-то современные общепринятые математические суждения (или суждения, которые мы будем считать неоспоримыми в будущем) содержат скрытые «врожденные» противоречия. Возможно, сошлется даже на тот знаменитый парадокс (о «множестве множеств, которые не являются элементами самих себя»), о котором Бертран Рассел писал Готтлобу Фреге в 1902 году, как раз тогда, когда Фреге собирался опубликовать труд всей своей жизни, посвященный основам математики (см. также комментарий к возражению **Q9**, § 2.7 и НРК, с. 100). В приложении к книге Фреге писал (см. [127]):

> Вряд ли с ученым может приключиться что-либо более нежеланное, чем потрясение основ его мировоззрения сразу вслед за тем, как он закончил изложение их на бумаге. Именно в такое положение поставило меня письмо от г-на Бертрана Рассела...



Разумеется, мы всегда можем сказать, что Фреге просто-напросто ошибся. Всем известно, что математики иногда допускают ошибки — порой даже весьма серьезные. Более того, как явствует из признания самого Фреге, его ошибка была вполне исправимой. Разве мы не убедились (в § 2.10, комментарий к **Q13**) в том, что подобные исправимые ошибки не имеют к нашим рассуждениям никакого отношения? Мы рассматриваем здесь, как и в § 2.10, лишь принципиальные вопросы, а не подверженность ошибкам отдельных представителей математического сообщества. Ошибки же, на которые можно указать, ошибочность которых можно однозначно продемонстрировать, вовсе не принадлежат к категории принципиальных вопросов, разве не так? Все так, однако ситуация, рассматриваемая нами в настоящий момент, несколько отличается от той, что обсуждалась в комментарии к возражению **Q13**, поскольку теперь у нас есть формальная система $\mathbb{F}$, которая, возможно, лежит в основе нашего математического понимания, только мы об этом не *знаем*. Как и прежде, нас не занимают единичные ошибки — или «оговорки», — которые может допустить отдельный математик, рассуждая в рамках какой-то в общем непротиворечивой системы. Однако теперь речь идет еще и о том, что сама система может содержать в себе некие глобальные противоречия. Именно это и произошло в случае с Фреге. Не узнай Фреге о парадоксе Рассела (или ином парадоксе сходной природы), вряд ли кто-либо смог бы убедить его в том, что в его систему вкралась фундаментальная ошибка. Дело не в том, что Рассел указал на какое-то формальное упущение в рассуждениях Фреге, а Фреге признал наличие ошибки, руководствуясь собственными канонами построения умозаключений; нет, Фреге продемонстрировали, что в самих этих канонах содержится некое изначальное противоречие. И именно факт наличия *противоречия*, а не что-либо иное, убедило Фреге в том, что его рассуждения ошибочны, а то, что прежде представлялось несокрушимой истиной, на деле фундаментально неверно. При этом о существовании ошибки стало известно только благодаря тому, что вскрылось противоречие. Если бы факт противоречивости установлен не был, то математики могли бы еще долгое время

считать предложенные Фреге методы построения умозаключений вполне достоверными и даже, возможно, строили бы на их фундаменте собственные системы.

Впрочем, полагаю, в данном случае крайне маловероятно, что многим математикам удалось бы в течение сколько-нибудь длительного срока наслаждаться той свободой умопостроений (в отношении бесконечных множеств), какую предоставляла система Фреге. Причина в том, что парадоксы типа парадокса Рассела довольно легко обнаружить. Можно представить себе какой-нибудь гораздо более тонкий парадокс, например, такой, что неявным образом содержится в тех или иных полагаемых нами на данный момент неопровержимо истинными математических процедурах, — парадокс, о котором никто не узнает еще, быть может, многие века. Необходимость в смене привычных правил мы осознáем лишь тогда, когда такой парадокс наконец себя проявит. Короче говоря, наша математическая интуиция не зиждется на каких-то непреходящих в веках установлениях, а напротив, непрерывно меняется под сильным воздействием идей, которые прекрасно «работали» *прежде*, и соображений, последствия применения которых пока что «сходят нам с рук». Такая точка зрения отнюдь не исключает возможности существования в основе нашего теперешнего математического понимания некоего алгоритма (или формальной системы), однако этот алгоритм не является чем-то неизменным, по мере обнаружения новых данных он подвергается непрерывной модификации. К изменяющимся алгоритмам мы еще вернемся несколько позднее (см. §§ 3.9–3.11, а также § 1.5), где и убедимся в том, что это по-прежнему все те же алгоритмы, только в ином обличье.

Разумеется, с моей стороны было бы наивным отрицать тот факт, что в методах, которые применяют в своей работе математики, нередко присутствует элемент «доверия» процедуре, если она «до сих пор, кажется, работает». В моей собственной математической практике такие предварительные, ориентировочные, нечеткие соображения составляют в общей совокупности рассуждений весьма заметный процент. Однако они, как правило, обретаются в той области, которая «отвечает» за нащупывание нового, еще не сформировавшегося понимания, а никак не в той, где мы «складываем» неопровержимо, на наш взгляд, установленные истины. Я очень сомневаюсь, что сам Фреге так уж категорически полагал свою систему абсолютно неопровержимой,



даже не подозревая еще о парадоксе, о котором написал ему Рассел. Система суждений столь общего характера, что бы ни думал по ее поводу автор, всегда выдвигается на всеобщее обозрение с некоторой настороженностью. Лишь после длительного «периода осмысления» можно будет полагать, что она достигла, наконец, «уровня неопровержимости». Имея же дело с системой настолько общей, как система Фреге, в любом случае, как мне кажется, следует употреблять выражения вида «полагая систему Фреге обоснованной, можно считать справедливым то-то и то-то», а не просто утверждать эти самые «то-то и то-то» без упомянутой оговорки. (См. также комментарии к возражениям **Q11** и **Q12**.)

Возможно, в настоящее время математики стали более осторожными в отношении того, что они готовы рассматривать как «неопровержимую истину» — эпоха осторожности сменила эпоху отчаянной дерзости (среди примеров которой работа Фреге занимает далеко не последнее место), пришедшуюся на конец XIX столетия. С выходом на сцену парадокса Рассела и прочих ему подобных необходимость в такой осторожности проявляется особенно наглядно. Что же касается дерзости, то она, по большей части, уходит корнями в те времена, когда математики начали потихоньку осознавать всю мощь канторовой теории бесконечных чисел и бесконечных множеств, выдвинутой им в начале того же XIX века. (Следует, впрочем, отметить, что Кантор знал о парадоксах, подобных парадоксу Рассела, — задолго до того, как сам Рассел обнаружил тот, что был назван его именем[4], — и предпринимал попытки усовершенствовать свою формулировку с тем, чтобы, по возможности, учитывать подобные проблемы.) Цели и характер моих рассуждений на этих страницах также, несомненно, требуют крайней осторожности. И я безмерно рад, что нам с вами приходится иметь дело только с утверждениями, истинность которых неопровержима, и что нет никакой необходимости влезать в дебри бесконечных множеств и прочих сомнительных понятий. Важно помнить, что — *где бы мы ни провели черту* — полученные с помощью доказательства Гёделя утверждения всегда остаются в рамках неопровержимо истинного (см. также комментарий к возражению **Q13**). Само по себе доказательство Гёделя (—Тьюринга) не имеет абсолютно никакого отношения к вопросам, связанным с сомнительным существованием бесконечных множеств определенного сорта. Неясности,

касающиеся тех самых исключительно вольных рассуждений, столь занимавших Кантора, Фреге и Рассела, ничуть не занимают *нас* — до тех пор, пока они остаются «сомнительными», не претендуя на звание «неопровержимых». Коль скоро мы со всем этим согласны, я никак не могу счесть правдоподобным допущение, согласно которому математики *действительно* используют в качестве основы для своего математического понимания и убеждений какую-либо необоснованную формальную систему $\mathbb{F}$. Я надеюсь, читатель согласится с тем, что вне зависимости от того, *возможна* такая ситуация или нет, она, во всяком случае, *невероятна*.

Наконец, в связи с возможной необоснованностью нашей гипотетической системы $\mathbb{F}$, вернемся ненадолго к другим аспектам человеческой «неточности», о которых мы говорили выше (см. комментарии к возражениям **Q12** и **Q13**). Прежде всего повторю: нас в данном случае интересуют *не* вдохновение, *не* гениальные догадки и *не* эвристические критерии, способные привести математика к великим открытиям, но лишь понимание и проникновение в суть, на фундаменте которых покоятся его неопровержимые убеждения в отношении математических истин. Эти убеждения могут оказаться всего-навсего результатом ознакомления с рассуждениями других математиков, и в этом случае о каких бы то ни было элементах математического открытия говорить, разумеется, не приходится. А вот когда мы нащупываем путь к какому-то подлинному открытию, и впрямь весьма важно дать размышлениям свободу, не ограничивая их изначально необходимостью в полной достоверности и точности (у меня сложилось впечатление, что именно это имел в виду Тьюринг в приведенной выше цитате, см. § 3.1). Однако когда перед нами встает вопрос о принятии или отклонении тех или иных доводов в поддержку неопровержимой истинности выдвигаемого математического утверждения, необходимо полагаться лишь на понимание и проницательность (нередко в сопровождении громоздких вычислений), которым ошибки принципиально не свойственны.

Я вовсе не хочу сказать, что математики, полагающиеся на понимание, не делают ошибок, — делают, и даже часто: понимание тоже можно применить некорректно. Безусловно, математики допускают ошибки и в рассуждениях, и в понимании, а также в сопутствующих вычислениях. Однако склонность к совершению подобных ошибок, в сущности, не *усиливает* их способности к



пониманию (хотя я, пожалуй, могу представить себе, каким образом подобные случайные обстоятельства могут порой привести человека к нежданному, скажем так, озарению). Что более важно — эти ошибки *исправимы*; их можно *распознать* как ошибки, когда на них укажет какой-либо другой математик (или даже впоследствии сам автор). Совсем иначе обстоит дело, когда понимание математика контролируется некоей внутренне ошибочной формальной системой $\mathbb{F}$: в рамках такой системы невозможно распознать ее собственные ошибки. (Что касается возможности существования самосовершенствующейся системы, которая модифицирует самое себя всякий раз, как обнаруживает в себе противоречие, то о ней мы поговорим несколько позднее, «на подступах» к противоречию § 3.14. Там же мы и обнаружим, что и от такого предположения в данном случае пользы мало; см. также § 3.26.)

Ошибки несколько иного рода возникают при неверной формулировке математического утверждения; в этом случае выдвигающий утверждение математик, возможно, *имеет в виду* нечто совсем отличное от того, что он буквально утверждает. Впрочем, такие ошибки также исправимы и не имеют ничего общего с теми *внутренними* ошибками, причиной которых является понимание, опирающееся на необоснованную систему $\mathbb{F}$ (здесь уместно вспомнить фразу Фейнмана, которую мы цитировали в связи с возражением **Q13**: «Не слушайте, что я говорю; слушайте, что я имею в виду!»). Мы с вами здесь для того, чтобы выяснить, что *в принципе* может (либо не может) быть установлено каким угодно математиком (человеком); ошибки же, подобные только что рассмотренным, — т. е. исправимые ошибки — никакого отношения к этой проблеме не имеют. Важнейший, пожалуй, для всего нашего исследования момент: круг идей и понятий, доступных математическому пониманию, непременно должен включать в себя центральную идею доказательства Гёделя—Тьюринга; на этом, собственно, основании мы и не рассматриваем всерьез возможность **I**, а возможность **II** полагаем крайне невероятной. Как уже отмечалось выше (в комментарии к возражению **Q13**), *идея* доказательства Гёделя—Тьюринга, безусловно, должна являться частью того, что *в принципе* в состоянии понять математик, даже если какое-то конкретное утверждение «$G(\mathbb{F})$», на котором этот математик, возможно, основывается, ошибочно — лишь бы ошибка была *исправимой*.

С возможной «необоснованностью» предполагаемого алгоритма математического понимания связаны и другие вопросы, о которых не следует забывать. Эти вопросы касаются процедур «восходящего» типа — таких, к примеру, как самоусовершенствующиеся алгоритмы, алгоритмы обучения (в том числе и искусственные нейронные сети), алгоритмы с дополнительными случайными компонентами, а также алгоритмы, операции которых обусловлены внешним окружением, в котором функционируют соответствующие алгоритмические устройства. Некоторые из упомянутых вопросов были затронуты ранее (см. комментарий к возражению **Q2**), подробнее же мы рассмотрим их при обсуждении случая **III**, к каковому обсуждению мы как раз и приступаем.

## 3.5. Может ли алгоритм быть непознаваемым?

В соответствии с вариантом **III**, математическое понимание представляет собой результат выполнения некоего непознаваемого алгоритма. Что же конкретно означает определение «непознаваемый» применительно к алгоритму? В предшествующих разделах настоящей главы мы занимались вопросами *принципиальными*. Так, утверждая, что неопровержимая истинность некоторого $\Pi_1$-высказывания доступна математическому пониманию человека, мы, по сути, утверждали, что данное $\Pi_1$-высказывание постижимо *в принципе*, отнюдь не имея в виду, что каждый математик когда-нибудь да сталкивался с реальной демонстрацией его истинности. Применительно к *алгоритму*, однако, нам потребуется несколько иная интерпретация термина «непознаваемый». Я буду понимать его так: рассматриваемый алгоритм является настолько сложным, что даже описание его *практически* неосуществимо.

Когда мы говорили о выводах, осуществляемых в рамках какой-то конкретной познаваемой формальной системы, или о предполагаемых результатах применения того или иного известного алгоритма, рассуждения в терминах принципиально возможного или невозможного и в самом деле выглядели как нельзя более уместными. Вопросы возможности или невозможности вывода того или иного конкретного предположения из такой формальной системы или алгоритма рассматривались в «принципиальном» контексте в силу элементарной *необходимости*. Похожим образом обстоит дело с установлением истинности $\Pi_1$-высказываний. $\Pi_1$-высказывание признается *истинным*, если его можно представить в виде операции некоторой машины Тьюринга, незавершаемой принципиально, вне зависимости от того, что мы могли бы получить на практике путем непосредственных вычислений. (Об этом мы говорили в комментарии к возражению **Q8**.) Аналогично, утверждение, что какое-то конкретное предположение выводимо (либо невыводимо) в рамках некоей формальной системы, следует понимать в «принципиальном» смысле, поскольку такое утверждение, в сущности, представляет собой вид утверждения об истинном (или, соответственно, ложном) характере какого-то конкретного $\Pi_1$-высказывания (см. окончание обсуждения возражения **Q10**). Соответственно, когда нас интересует выводимость предположения в рамках некоторого неизменного набора правил, «познаваемость» всегда будет пониматься именно в таком «принципиальном» смысле.



Если же нам предстоит решить вопрос о «познаваемости» самих правил, то здесь необходимо прибегнуть к «практическому» подходу. *Принципиально* возможно описать *любую* формальную систему, машину Тьюринга, либо $\Pi_1$-высказывание, а следовательно, если мы хотим, чтобы вопрос об их «непознаваемости» имел хоть какой-нибудь смысл, нам следует рассматривать его именно в плоскости возможности их практической реализации. В принципе, познаваемым является абсолютно любой алгоритм, каким бы он ни был, — в том смысле, что осуществляющая этот алгоритм операция машины Тьюринга становится «известной», как только становится известным натуральное число, являющееся кодовым обозначением данной операции (например, согласно правилам нумерации машин Тьюринга, приведенным в НРК). Нет решительно никаких оснований предполагать, что принципиально непознаваемым может оказаться такой объект, как натуральное число. Все натуральные числа (а значит, и алгоритмические операции) можно представить в виде последовательности 0, 1, 2, 3, 4, 5, 6, ..., двигаясь вдоль которой, мы — *в принципе* — можем со временем достичь любого натурального числа, каким бы большим это число ни было! Практически же, число может оказаться настолько огромным, что добраться до него таким способом в обозримом будущем не представляется возможным. Например, номер машины Тьюринга, описанной в НРК (на с. 56), явно слишком велик, чтобы его можно было получить на практике посредством подобного перечисления.

Даже если мы были бы способны выдавать каждую последующую цифру за наименьший теоретически определимый временной промежуток (в масштабе времени Планка равный приблизительно $0{,}5 \times 10^{-43}$ с, см. §6.11), то и в этом случае за все время существования Вселенной, начиная от Большого Взрыва и до настоящего момента, нам не удалось бы добраться до числа, двоичное представление которого содержит более 203 знаков. В числе, о котором только что упоминалось, знаков более чем в 20 раз больше — однако это ничуть не мешает ему быть «познаваемым» в принципе, причем в НРК это число определено в явном виде.

Практически «непознаваемым» следует считать такое натуральное число (или операцию машины Тьюринга), сложность одного только описания которого оказывается недоступной человеческим возможностям. Сказано, на первый взгляд, довольно громко, однако, зная о конечной природе человека, можно смело утверждать, что *какой-то* предел так или иначе существовать должен, а следовательно, должны существовать и числа, находящиеся *за* этим пределом, описать которые человек не в состоянии. (См. также комментарий к возражению **Q8**.) В соответствии с возможностью **III**, нам следует полагать, что за пределами познаваемости алгоритм $F$ (предположительно лежащий в основе математического понимания) оказывается именно вследствие неимоверной сложности и чрезвычайной детализированности своего описания — причем речь идет исключительно об «*описуемости*» алгоритма, а не о познаваемости его как алгоритма, которым, предполагается, мы пользуемся-таки в нашей интеллектуальной деятельности. Требование «неописуемости», собственно, и отделяет случай **III** от случая **II**. Иными словами, рассматривая случай **III**, мы должны учитывать возможность того, что наших человеческих способностей может оказаться недостаточно даже для того, чтобы описать это самое число, не говоря уже о том, чтобы установить, обладает ли оно свойствами, какими должно обладать число, определяющее алгоритмическую операцию, в соответствии с которой работает наше же математическое понимание.

Отметим, что в роли ограничителя познаваемости не может выступать просто величина числа. Не представляет никакой сложности описать числа, настолько огромные, что они *превзойдут* по величине все числа, которые могут потребоваться для



описания алгоритмических операций, определяющих поведение любого организма в наблюдаемой Вселенной (взять хотя бы такое легко описываемое число, как $2^{2^{65536}}$, о котором мы упоминали в комментарии к **Q8**, — это число далеко превосходит количество всех возможных состояний Вселенной для всего вещества, содержащегося в границах наблюдаемой нами Вселенной[5]). За пределами человеческих возможностей должно оказаться именно *точное* описание искомого числа, величина же его особой роли не играет.

Допустим (в полном согласии с **III**), что описание такого алгоритма $F$ человеку и в самом деле не по силам. Что из этого следует в отношении перспектив разработки высокоуспешной стратегии создания ИИ (как по «сильным», так и по «слабым» принципам — иначе говоря, в соответствии с точками зрения как $\mathscr{A}$, так и $\mathscr{B}$)? Адепты полностью автоматизированных ИИ-систем (т. е. сторонники $\mathscr{A}$ непременно, а также, возможно, кто-то из лагеря $\mathscr{B}$) предвосхищают появление в конечном итоге роботов, способных достичь уровня математических способностей человека и, возможно, превзойти этот уровень. Иными словами (если согласиться с вариантом **III**), непременным компонентом контрольной системы такого робота-математика должен стать тот самый, недоступный человеческому пониманию алгоритм $F$. Отсюда, по всей видимости, следует, что стратегия создания ИИ, нацеленная на получение именно такого результата, обречена на провал. Причина проста — если для достижения цели необходим алгоритм $F$, который в принципе не способен описать ни один человек, то где же тогда этот алгоритм взять?

Однако наиболее амбициозные сторонники идеи ИИ рисуют себе совсем другие картины. Они предвидят, что необходимый алгоритм $F$ будет получен не в одночасье, но поэтапно — по мере того, как сами роботы будут постепенно повышать свою эффективность с помощью алгоритмов (восходящих) обучения и накопления опыта. Более того, самые совершенные роботы не будут, скорее всего, созданы непосредственно людьми, а явятся продуктом деятельности других роботов[6], возможно, несколько более примитивных, нежели ожидаемые нами роботы-математики; кроме того, в процессе развития роботов будет, возможно, принимать участие и некое подобие дарвиновской эволюции, в результате чего от поколения к поколению роботы будут становиться все более совершенными. Разумеется, не обходится и без утверждений

в том духе, что именно посредством подобных, в общем-то, процессов нам самим удалось оснастить свои «нейронные компьютеры» неким для нас не познаваемым алгоритмом $F$, на котором и работает наше собственное математическое понимание.

В нескольких последующих разделах я покажу, что при всей привлекательности подобных процессов проблема, в сущности, остается нерешенной: если сами процедуры, с помощью которых предполагается создать ИИ, являются прежде всего алгоритмическими и познаваемыми, то любой полученный таким образом алгоритм $F$ также должен быть познаваемым. В этом случае вариант **III** сводится либо к варианту **I**, либо к варианту **II**, которые мы исключили в §§ 3.2–3.4 по причине фактической невозможности (вариант **I**) или, по меньшей мере, крайнего неправдоподобия (вариант **II**). Более того, если исходить из допущения, что интересующие нас алгоритмические процедуры познаваемы, то нам, вообще говоря, следует отдать предпочтение именно варианту **I**. Соответственно, вариант **III** (равно как и, по смыслу, вариант **II**) также следует признать практически несостоятельным.

Читателю, который искренне верит в то, что возможный вариант **III** открывает наиболее вероятный путь к созданию вычислительной модели разума, я рекомендую обратить на приведенные выше аргументы самое пристальное внимание и тщательнейшим образом их изучить. Не сомневаюсь, что он придет к тому же выводу, к какому пришел я: если допустить, что математическое понимание и в самом деле осуществляется в соответствии с вариантом **III**, то единственным хоть сколько-нибудь правдоподобным объяснением происхождения нашего собственного алгоритма $F$ остается считать божественное вмешательство — то самое сочетание $\mathscr{A}/\mathscr{D}$, о котором мы говорили в конце § 1.3, — а такое объяснение, конечно же, не утешит тех, кто лелеет амбициозные перспективные планы по созданию компьютерного ИИ.

## 3.6. Естественный отбор или промысел Господень?

Возможно, нам следует-таки всерьез рассмотреть возможность того, что за нашим интеллектом и в самом деле стоит некий божественный промысел — по каковой причине этот самый интеллект никак нельзя объяснить с позиций той науки, которая



достигла столь значительных успехов в описании мира неодушевленных предметов. Разумеется, мы по-прежнему будем сохранять широту мышления, однако я хочу сразу прояснить один момент: в последующих рассуждениях я намерен придерживаться научной точки зрения. Я намерен рассмотреть возможность того, что наше математическое понимание является результатом работы некоего непостижимого алгоритма, — а также вопрос о возможном происхождении подобного алгоритма, — никоим образом не выходя за рамки научного подхода. Возможно, кто-то из читателей этой книги склонен верить в то, что этот алгоритм и в самом деле мог быть просто вложен в наши головы по воле божьей. Убедительного опровержения такого предположения у меня, признаться, нет; хотя я никак не могу взять в толк, — если уж мы решаем отказаться на каком-то этапе от научного подхода — почему считается как нельзя более благоразумным бросаться именно в эту крайность. Если научное объяснение ничего, в сущности, не объясняет, то не уместнее ли будет вообще позабыть о каких бы то ни было алгоритмических процедурах, нежели прятать свою предполагаемую свободу воли за сложностью и непостижимостью какого-то алгоритма, который, как нам хочется думать, контролирует каждое наше движение? Возможно, разумнее будет просто счесть (как, похоже, считал сам Гёдель), что деятельность разума совершенно не связана с процессами, протекающими в физическом мозге, — что замечательно согласуется с точкой зрения $\mathscr{D}$. С другой стороны, в настоящее время, как мне представляется, даже те, кто верит в то, что мышление и впрямь является в каком-то смысле божественным даром, склонны все же полагать, что поведение человека можно объяснить, не выходя за пределы возможностей науки. Несомненно, приведенные варианты являются весьма спорными, однако на данном этапе я вовсе не предполагал спорить с убеждениями сторонников точки зрения $\mathscr{D}$. Надеюсь, что те читатели, которых можно отнести к приверженцам той или иной формы $\mathscr{D}$, все же потерпят меня еще некоторое время, а я пока попробую выяснить, к чему нас может привести в данном случае научный подход.

Какие же научные последствия может иметь допущение, что математические суждения мы получаем в результате выполнения некоей необходимой и непостижимой алгоритмической процедуры? Вырисовывается приблизительно такая картина: исключительно сложные алгоритмические процедуры, необходимые для

моделирования подлинного математического понимания, являются результатом многих сотен тысяч лет (по меньшей мере) естественного отбора вкупе с несколькими тысячами лет воздействия обучения и внешних факторов, обусловленных физическим окружением. Можно допустить, что наследуемые аспекты этих процедур формировались постепенно из более простых (ранних) алгоритмических компонентов в результате того же давления естественного отбора, которое ответственно за возникновение всех остальных в высшей степени эффективных механизмов, из которых составлены как наши тела, так и наши мозги. Врожденные потенциально математические алгоритмы (т. е. все те унаследованные аспекты, которые могли бы относиться к математическому мышлению, предположительно алгоритмическому) до поры пребывали в закодированном состоянии (в виде неких особых последовательностей нуклеотидов) внутри молекул ДНК, а затем проявились посредством той же процедуры, какая задействуется при всяком постепенном (либо скачкообразном) усовершенствовании живого организма, реагирующего на давление отбора. Помимо прочего, свой вклад в эти процессы вносят и всевозможные внешние факторы — такие как непосредственное математическое образование, опыт взаимодействия с физическим окружением, прочие факторы, оказывающие дополнительно самые разные чисто случайные воздействия. Думаю, мы должны попытаться выяснить, можно ли полагать описанную картину хоть сколько-нибудь правдоподобной?

## 3.7. Алгоритм или алгоритмы?

Прежде всего, необходимо рассмотреть следующий весьма важный вопрос: может ли оказаться, что за различные виды математического понимания, свойственные разным людям, отвечает множество весьма различных, возможно, неэквивалентных алгоритмов? В самом деле, уж в чем мы можем быть с самого начала уверены, так это в том, что даже профессиональные математики часто воспринимают математические «реалии» совершенно поразному. Для одних в высшей степени важны зрительные образы, тогда как другим удобнее иметь дело с четкими логическими структурами, изящными абстрактными доказательствами, подробными аналитическими обоснованиями или, возможно, чисто алгебраическими манипуляциями. В этой связи следует отметить,



что, по некоторым предположениям, геометрическое, например, и аналитическое мышление осуществляются разными полушариями мозга (соответственно, правым и левым)[7]. Однако часто бывает так, что всеми этими способами воспринимается одна и та же математическая истина. С алгоритмической точки зрения первое впечатление таково: алгоритмы, отвечающие за математическое мышление различных людей, должны быть как минимум абсолютно неэквивалентными. Однако, несмотря на существенное различие между образами, которые формируют в сознании отдельные математики (или прочие смертные) для собственного понимания или для сообщения другим математических идей, математическое восприятие обладает одним поразительным свойством: когда математики наконец решают для себя, что именно следует считать неопровержимо истинным, никаких разногласий по этому поводу больше не возникает, разве что поводом для такого разногласия послужит какая-либо действительная, опознаваемая (а следовательно, и исправимая) ошибка в рассуждениях того или иного математика (еще один возможный повод для разногласий предоставляет принципиальное расхождение во мнениях по некоторым — весьма немногочисленным — фундаментальным вопросам; см. комментарий к **Q11**, в особенности утверждение $\mathcal{G}^{***}$). В целях упрощения изложения я позволю себе в дальнейшем последнее соображение проигнорировать. Хотя это соображение и имеет некоторое отношение к предмету нашего разговора, на выводы оно заметного влияния не оказывает. (Придерживаемся ли мы нескольких возможных неэквивалентных точек зрения на какой-то вопрос или все соглашаемся на одной — существенного различия между этими двумя ситуациями в данном случае нет.)

Восприятие математической истины может осуществляться самыми различными способами. Вряд ли можно усомниться в том, что вне зависимости от конкретной природы физических процессов, обусловливающих осознание человеком истинности какого-либо математического утверждения, эти процессы должны весьма и весьма разниться от индивидуума к индивидууму, даже если речь идет об одном и том же утверждении. Иначе говоря, если математики при составлении суждений о неопровержимой истинности того или иного утверждения просто-напросто применяют какие-то вычислительные алгоритмы, то у разных математиков эти самые алгоритмы должны весьма значительно

различаться по своей структуре. При этом упомянутые алгоритмы должны быть еще и *эквивалентны* друг другу в некотором очевидном смысле.

Это условие, возможно, не так уж и абсурдно, как может показаться на первый взгляд — по крайней мере, с точки зрения математически *возможного*. Весьма разные на вид машины Тьюринга могут давать на выходе идентичные результаты. (Рассмотрим, например, машину Тьюринга, построенную следующим образом: при выполнении действия над натуральным числом $n$ мы получаем в результате 0 всякий раз, когда $n$ выразимо в виде суммы четырех квадратов, и 1, когда $n$ таким образом выразить нельзя. Результат вычисления такой машины полностью совпадает с результатом другой машины, построенной таким образом, чтобы давать на выходе 0 при подаче на вход *любого* натурального числа $n$ — ибо известно, что в виде суммы четырех квадратов можно представить *любое* натуральное число; см. § 2.3.) Из идентичности внешних конечных результатов двух алгоритмов вовсе не обязательно следует, что эти алгоритмы окажутся подобными по внутренней структуре. Однако, в определенном смысле, рассматриваемое допущение еще *более* запутывает вопрос о происхождении нашего гипотетического непостижимого алгоритма(-ов) для установления математической истины, поскольку теперь нам предстоит иметь дело уже с несколькими такими алгоритмами, достаточно отличными друг от друга по внутренней структуре, но при этом существенно эквивалентными в отношении получаемого на выходе результата.

## 3.8. Эзотерические математики не от мира сего как результат естественного отбора

Какую же роль играет во всем этом естественный отбор? Возможно ли, чтобы естественным путем возник некий алгоритм $F$ (или несколько таких алгоритмов), обусловливающий наше математическое понимание и при этом непознаваемый сам по себе (если верить допущению **III**), либо лишь в отношении выполняемых им функций (в соответствии с допущением **II**)? Начнем с повторения того, о чем мы уже говорили в начале § 3.1. В процессе получения своих предположительно неопровержимо истинных математических выводов математики *вовсе не считают*, что они



просто следуют некоему набору непознаваемых правил — правил настолько сложных, что, с математической точки зрения, они непостижимы в принципе. Напротив, они полагают, что эти выводы представляют собой результат неких обоснованных рассуждений (пусть зачастую длинных и внешне запутанных), которые в конечном счете опираются на четкие неопровержимые истины, понятные, в принципе, любому.

Более того, рассматривая ситуацию с позиций здравого смысла или на уровне логических дескрипций, мы можем со всей определенностью утверждать, что математики *и в самом деле* делают то, что, как им кажется, они делают. Этот факт не подлежит никакому сомнению, а важность его переоценить невозможно. Если мы полагаем, что математики в своей деятельности следуют некоему набору непознаваемых и непостижимых вычислительных правил (в соответствии с возможными вариантами **III** или **II**), то, значит, они делают *еще* и это — одновременно с тем, что, как им кажется, они делают, но на другом уровне дескрипции. Каким-то образом алгоритмическое следование правилам должно давать тот же самый *результат*, что дают математическое понимание и интуиция — по крайней мере, на практике. Если уж мы твердо вознамерились стать приверженцами либо $\mathcal{A}$, либо $\mathcal{D}$, то нам предстоит попытаться поверить в то, что такая возможность является вполне правдоподобной.

Нужно помнить и о том, какие блага дают эти алгоритмы. Предполагается, что они наделяют своего «носителя» — по крайней мере, в принципе — способностью составлять корректные математические суждения об абстрактных сущностях, весьма далеких от непосредственного жизненного опыта, что, по большей части, не дает этому самому носителю сколько-нибудь заметных практических преимуществ. Любой, кому хоть раз доводилось заглянуть в какой-нибудь современный чисто математический научный журнал, знает, насколько далеки заботы математиков от каких бы то ни было практических вопросов. Тонкости теоретических обоснований, обычно публикуемых в таких научных журналах, непосредственно доступны лишь очень небольшому количеству людей; и все же каждое такое рассуждение состоит, в конечном счете, из каких-то элементарных шагов, и каждый такой шаг может, *в принципе*, понять любой мыслящий индивидуум, даже если речь идет об абстрактных рассуждениях о сложно определяемых бесконечных множествах. Не следует забывать и

о том, что алгоритм — или, возможно, целый ряд альтернативных, но математически эквивалентных алгоритмов, — который дает человеку потенциальную способность понимать упомянутые рассуждения, каким-то образом был изначально записан не где-нибудь, а в нуклеотидных последовательностях молекулы ДНК. Если мы в это верим, то нам следует весьма серьезно задуматься, как же так получилось, что подобный алгоритм (или алгоритмы) развился в результате естественного отбора. Очевидно, что даже в настоящее время профессия математика не дает никаких преимуществ с точки зрения борьбы за существование. (Подозреваю, что ее можно даже считать неблагоприятным фактором. Вследствие своего взрывного темперамента и странноватых пристрастий пуристы со склонностью к математике имеют тенденцию заканчивать свой жизненный путь на какой-нибудь низкооплачиваемой академической службе — или и вовсе безработными.) Гораздо правдоподобнее выглядит иная картина: способность рассуждать о весьма абстрактно определяемых бесконечных множествах, бесконечных множествах бесконечных множеств и т. д. никаких особых преимуществ в борьбе за выживание нашим далеким предкам дать просто не могла. Этих самых предков заботили практические повседневные проблемы — такие, как постройка убежищ, изготовление одежды, изобретение ловушки для мамонтов или, несколько позднее, одомашнивание животных и выращивание урожая (см. рис. 3.1).

Разумно предположить, что упомянутые преимущества, которыми, очевидно, все же обладали наши предки, происходили из качеств, необходимых для решения как раз таких, практических проблем, а уже потом, гораздо позднее, выяснилось, что эти же качества замечательно подходят и для решения проблем математических — этакий *побочный* результат. Во всяком случае, такой ход событий полагаю более или менее правдоподобным я сам. Развивая это предположение, можно допустить, что под давлением естественного отбора человек каким-то образом приобрел или развил в себе некую общую способность *понимать*. Эта способность понимать, проникать в суть вещей, не была связана с какими-то конкретными областями его деятельности и оказывалась полезной буквально во всем. То же сооружение жилищ или ловушек для мамонтов существенно усложнилось бы, не обладай человек способностью понимать вещи и явления в их общности. При этом лично я полагаю, что *Homo sapiens* был



Рис. 3.1. Вряд ли специфическая способность составлять сложные математические суждения могла дать нашим далеким предкам какие бы то ни было преимущества в борьбе за существование, а вот общая способность к *пониманию* им наверняка не помешала бы.

отнюдь не уникален в своей способности понимать. Такой же способностью обладали, возможно, и многие другие животные, составлявшие человеку конкуренцию в борьбе за существование, однако обладали в меньшей степени, в результате чего человек, в силу более *интенсивного* развития этой способности, получил над ними весьма существенное преимущество.

Сложности с такой точкой зрения возникают как раз тогда, когда мы начинаем рассматривать наследуемую способность к пониманию как нечто по своей природе алгоритмическое. Как нам уже известно из предшествующих рассуждений и доказательств, любая (алгоритмическая) способность к пониманию, достаточно сильная для того, чтобы ее обладатель оказался в состоянии разобраться в тонкостях математических обоснований, в частности, гёделевского доказательства в представленном мною варианте, должна быть обусловлена процедурой настолько замысловатой и непостижимой, что о ней (или ее роли) не может знать даже сам обладатель этой способности. Наш прошедший через испытания естественного отбора гипотетический алгоритм, по всей видимости, достаточно силен, ведь еще во времена на-

ших далеких предков он уже включал в область своей потенциальной применимости правила всех формальных систем, рассматриваемых сегодня математиками как безоговорочно непротиворечивые (или неопровержимо обоснованные, если речь идет о $\Pi_1$-высказываниях, см. §2.10, комментарий к **Q10**). Сюда почти наверняка входят и правила формальной системы Цермело–Френкеля ZF, или, возможно, ее расширенного варианта, системы ZFC (иначе говоря, самой ZF с добавлением аксиомы выбора) — системы (см. §§3.3 и 2.10, комментарий к **Q10**), которую многие математики сегодня рассматривают как источник абсолютно всех необходимых для обычной математики методов построения рассуждений, — а также все частные формальные системы, получаемые из системы ZF посредством применения к ней процедуры гёделизации сколько угодно раз, и кроме того, все другие формальные системы, которые могут быть получены математиками посредством тех или иных озарений и рассуждений — скажем, на основании открытия, суть которого состоит в том, что системы, полученные в результате упомянутой гёделизации, всегда являются неопровержимо обоснованными, или исходя из иных рассуждений еще более основополагающего характера. Такой алгоритм должен был также включать в себя (в виде собственных частных экземпляров) потенциальные способности к установлению тонких различий, отделению справедливых аргументов от ничем не обоснованных во всех тех, тогда еще не открытых, областях математики, которые сегодня оккупируют страницы специальных научных журналов. Все вышеперечисленные способности должны были оказаться каким-то образом закодированы внутри этого самого — гипотетического, непознаваемого или, если угодно, непостижимого — алгоритма, и вы хотите, чтобы мы поверили, что он возник исключительно в результате естественного отбора, в ответ на какие-то внешние условия, в которых нашим далеким предкам приходилось бороться за выживание. Конкретная способность к отвлеченным математическим рассуждениям не могла дать своему обладателю никаких непосредственных преимуществ в этой борьбе, и я со всей определенностью утверждаю, что для возникновения подобного алгоритма не существовало и не могло существовать никаких естественных причин.

Однако стоит нам допустить, что «способность понимать» имеет неалгоритмическую природу, как ситуация в корне меняется. Теперь уже нет необходимости приписывать этой способности какую-то неимоверную сложность, вплоть до полной непознаваемости или непостижимости. Более того, она может оказаться гораздо ближе к тому, что «математики, как им кажется, делают». Способность к пониманию представляется мне весьма простым и даже обыденным качеством. Ее сложно определить в каких-либо точных терминах, однако она настолько близка нам и привычна, что в принципиальную невозможность корректного моделирования понимания посредством какой бы то ни было вычислительной процедуры верится с трудом. И все же так оно и есть. Для создания подобной вычислительной модели необходима алгоритмическая процедура, так или иначе учитывающая все возможные варианты развития событий в будущем, — т. е. алгоритм, в котором должны быть, скажем так, предварительно запрограммированы ответы на все математические вопросы, с которыми нам когда-либо предстоит столкнуться. Если непосредственному программированию эти ответы не подлежат, то нужно обеспечить какие-то вычислительные способы для их отыскания. Как мы уже успели убедиться, если эти «вычислительные способы» (или «предварительное программирование») охватывают все, что когда-либо было или будет доступно человеческому пониманию, то сами они для человека становятся непостижимыми. Откуда же слепым эволюционным процессам, нацеленным исключительно на обеспечение выживания сильнейших, было «знать» о том, что такая-то непознаваемо обоснованная вычислительная процедура окажется когда-то в будущем способной решать абстрактные математические задачи, не имеющие абсолютно никакого отношения к проблемам выживания?



## 3.9. Алгоритмы обучения

Дабы не подвергать читателя искушению чересчур поспешно смириться с абсурдностью описанной выше возможности, я должен несколько прояснить картину, на что мне уже, несомненно, указывают сторонники вычислительного подхода. Как уже отмечалось в §3.5, эти самые сторонники имеют в виду не столько алгоритм, который, в известном смысле, «предварительно запрограммирован» на предоставление решений математических проблем, сколько некую вычислительную систему, способную *обу-*

*чаться*. Такая система может состоять, в основе своей, из «восходящих» компонентов, соединенных по мере необходимости с какими-либо «нисходящими» процедурами (см. § 1.5)[4].

Возможно, кому-то покажется, что называть «нисходящей» систему, возникшую исключительно в результате слепого давления естественного отбора, не совсем уместно. Этим термином я буду обозначать здесь те аспекты нашей гипотетической алгоритмической процедуры, которые для данного организма *зафиксированы* генетически и не подвержены изменению под влиянием последующего жизненного опыта или обучения каждого отдельного представителя вида. Хотя упомянутые нисходящие аспекты и не были созданы кем-то или чем-то, обладающим подлинным «знанием» об их предполагаемых функциях и возможностях (речь идет всего лишь о трансляции определенных цепочек ДНК, приводящей к соответствующей активности клеток мозга), они, тем не менее, способны четко обозначить правила, в соответствии с которыми и будет действовать математически активный мозг. Эти нисходящие процедуры снабдят нашу систему теми алгоритмическими операциями, которые составят необходимую фиксированную структуру, в рамках которой, в свою очередь, будут функционировать более гибкие «процедуры обучения» (восходящие).

Какова же природа этих процедур обучения? Вообразим, что наша самообучающаяся система помещена в некоторое внешнее окружение, причем поведение системы внутри этого окружения непрерывно модифицируется под влиянием реакции окружения на ее предыдущие действия. В процессе участвуют, в основном, два фактора. ***Внешним*** фактором является поведение окружения и его реакция на действия системы, а ***внутренним*** — изменения в поведении системы в ответ на изменения в окружении. Прежде всего следует решить вопрос об алгоритмической природе внешнего фактора. Мо-

[4]На сегодняшний день мы располагаем вполне строгой математической теорией обучения; см. [10]. Однако эта теория имеет отношение больше к *сложности*, нежели к вычислимости — иными словами, рассматривает вопросы, связанные с производительностью вычислительных машин и объемом их памяти, необходимыми для решения тех или иных проблем; см. НРК, с. 140–145. Создатели теории не делают никаких предположений о том, что такие математически определенные системы обучения могут оказаться способными моделировать процесс приобретения математиком-человеком собственного понятия о «неопровержимой истине».



жет ли реакция внешнего окружения вносить в общую картину некую неалгоритмическую составляющую, если внутреннее устройство нашей системы обучения является целиком и полностью алгоритмическим?

В определенных обстоятельствах (как, например, часто бывает при «обучении» искусственных нейронных сетей) реакция внешнего окружения заключается в изменении поведения экспериментатора (инструктора, преподавателя — в дальнейшем предлагаю называть его просто «учителем»), изменении намеренном и предпринимаемом с целью улучшить качество функционирования системы. Когда система функционирует так, как требует учитель, ей об этом сообщают, чтобы в дальнейшем (под воздействием внутренних механизмов модификации поведения системы) она с большей вероятностью функционировала бы именно таким образом. Предположим, например, что у нас имеется искусственная нейронная сеть, которую необходимо научить распознавать человеческие лица. Мы непрерывно наблюдаем за функционированием нашей системы и после каждого рабочего цикла снабжаем ее данными о правильности ее последних «догадок» для того, чтобы она могла улучшить качество своей работы, модифицировав нужным образом внутреннюю структуру. На практике, за адекватностью результатов каждого рабочего цикла совсем не обязательно должен наблюдать учитель-человек, так как процедуру обучения можно в значительной степени автоматизировать. В описанной ситуации цели и суждения учителя-человека образуют наивысший критерий качества функционирования системы. В других ситуациях реакция окружения может оказаться не столь «преднамеренной». Например, в процессе развития *живых* систем — предполагается, что эти системы все же функционируют в соответствии с некоторой нейронной схемой (или иной алгоритмической процедурой, например, генетическим алгоритмом, см. § 3.7), вроде тех, что применяются в численном моделировании — в подобных внешних целях или суждениях вообще не возникает необходимости. Вместо этого, живые системы модифицируют свое поведение в процессе, который можно рассматривать как своего рода *естественный отбор*, действуя согласно критериям, эволюционировавшим на протяжении многих лет и способствующим увеличению шансов на выживание как самой системы, так и ее потомства.

## 3.10. Может ли окружение вносить неалгоритмический внешний фактор?

Выше мы предположили, что сама наша система (независимо от того, живая она или нет) представляет собой нечто вроде *робота* с компьютерным управлением, т. е. все ее самомодификационные процедуры являются целиком вычислительными. (Я пользуюсь здесь термином «робот» исключительно для того, чтобы подчеркнуть то обстоятельство, что нашу систему следует рассматривать как некую самостоятельную, целиком и полностью вычислительную сущность, находящуюся во взаимодействии со своим окружением. Я вовсе не подразумеваю, что она непременно представляет собой какое бы то ни было механическое устройство, целенаправленно сконструированное человеком. Такой системой, если верить $\mathscr{A}$ или $\mathscr{B}$, может оказаться развивающееся человеческое существо, а может и в самом деле какой-то искусственно созданный объект.) Итак, мы полагаем, что *внутренний* фактор является полностью вычислительным. Необходимо установить, является ли вычислительным также и *внешний* фактор, вносимый окружением, — иначе говоря, возможно ли построить эффективную численную модель этого самого окружения как в *искусственном* (т. е. когда окружение неким искусственным образом контролируется учителем-человеком), так и в *естественном* случае (когда высшим авторитетом является давление естественного отбора). В каждом случае конкретные внутренние правила, в соответствии с которыми система обучения робота модифицирует его поведение, должны быть составлены так, чтобы тем или иным образом реагировать на конкретные сигналы, посредством которых окружение будет сообщать системе о том, как следует оценивать качество ее функционирования в предыдущем рабочем цикле.

Вопрос о возможности моделирования окружения в искусственном случае (иными словами, о возможности численного моделирования поведения человека-учителя) представляет собой тот самый общий вопрос, ответ на который мы пытаемся найти вот уже в который раз. В рамках гипотез $\mathscr{A}$ или $\mathscr{B}$, следствия из которых мы рассматриваем в настоящий момент, допускается, что эффективное моделирование в этом случае и в самом деле возможно, по крайней мере, в принципе. В конце концов, цель нашего исследования состоит именно в выяснении общего правдоподобия этого допущения. Поэтому, вместе с допущением о вычислительной природе нашего робота, допустим также, что его окружение также вычислимо. В результате мы получаем *объединенную* систему, состоящую из робота и его обучающего окружения, которая, в принципе, допускает эффективное численное моделирование, т. е. окружение не дает никаких потенциальных оправданий невычислительному поведению вычислительного робота.



Иногда можно услышать утверждение, что нашим преимуществом перед компьютерами мы обязаны тому факту, что люди образуют *сообщество*, внутри которого происходит непрерывное общение между индивидуумами. Согласно этому утверждению, отдельного человека можно рассматривать как вычислительную систему, тогда как сообщество людей представляет собой уже нечто большее. То же относится и, в частности, к математическому сообществу и отдельным математикам — сообщество может вести себя невычислительным образом, в то время как отдельные математики такой способностью не обладают. На мой взгляд, это утверждение лишено всякого смысла. В самом деле, представьте себе аналогичное сообщество непрерывно общающихся между собой компьютеров. Подобное «сообщество» в целом является точно такой же вычислительной системой; деятельность его, если есть такое желание, можно смоделировать и на одном-единственном компьютере. Разумеется, вследствие одного только количественного превосходства, сообщество составит гораздо более мощную вычислительную систему, нежели каждый из индивидуумов в отдельности, однако *принципиальной* разницы между ними нет. Известно, что на нашей планете проживает более $5 \times 10^9$ человек (прибавьте к этому еще огромные библиотеки накопленного знания). Цифры впечатляют, но это всего лишь цифры — если отдельного человека считать вычислительным устройством, то разницу, обусловленную переходом от индивидуума к сообществу, развитие компьютерных технологий сможет при необходимости свести на нет в течение каких-нибудь нескольких десятилетий. Очевидно, что искусственный случай с учителями-людьми в роли внешнего окружения не дает нам ничего принципиально нового, что могло бы объяснить, каким образом из целиком и полностью вычислительных составляющих возникает абсолютно невычислимая сущность.

Что же мы имеем в естественном случае? Вопрос теперь звучит так: может ли физическое окружение (если не учитывать действий присутствующих в нем учителей-людей) содержать компоненты, которые невозможно даже в принципе смоделировать численными методами? Мне думается, что если кто-то полагает, что в «бесчеловечном» окружении может присутствовать нечто, принципиально не поддающееся численному моделированию, то этот кто-то тем самым лишает силы главное возражение против $\mathscr{C}$. Ибо единственной разумной причиной усомниться в возможной справедливости точки зрения $\mathscr{C}$ можно счесть лишь скептическое отношение к утверждению, что объекты, принадлежащие реальному физическому миру могут вести себя каким-то невычислимым образом. Как только мы признаём, что *какой-либо* физический процесс может оказаться невычислимым, у нас не остается никакого права отказывать в невычислимости и процессам, протекающим внутри такого физического объекта, как мозг, — равно как и возражать против $\mathscr{C}$. Как бы то ни было, крайне маловероятно, что в безлюдном окружении может обнаружиться нечто такое, что не поддается вычислению столь же фундаментально, как это делают некоторые процессы внутри человеческого тела. (См. также §§ 1.9 и 2.6, **Q2**.) Думаю, мало кто всерьез полагает, что среди всего, что имеет хоть какое-то отношение к окружению самообучающегося робота, может оказаться что-либо, *принципиально* невычислимое.

Впрочем, говоря о «принципиально» вычислимой природе окружения, не следует забывать об одном важном моменте. Вне всякого сомнения, на *реальное* окружение любого развивающегося живого организма (или некоей изощренной робототехнической системы) оказывают влияние весьма многочисленные и порой невероятно сложные факторы, вследствие чего любое моделирование этого окружения со сколько-нибудь приемлемой точностью вполне может оказаться неосуществимым *практически*. Динамическое поведение даже относительно простых физических систем бывает порой чрезвычайно сложным, при этом его зависимость от мельчайших нюансов начального состояния может быть настолько критической, что предсказать дальнейшее поведение такой системы решительно невозможно — в качестве примера можно привести ставшую уже притчей во языцех проблему долгосрочного предсказания погоды. Подобные системы называют *хаотическими*; см. § 1.7. (Хаотические системы характеризуются сложным и эффективно непредсказуемым поведением. Однако математически эти системы объяснить вполне возможно; более того, их активное изучение составляет весьма существенную долю современных математических исследований[8].) Как уже указывалось в § 1.7, хаотические системы я *также* включаю в категорию «вычислительных» (или «алгоритмических»). Для наших целей важно подчеркнуть один существенный момент, касающийся хаотических систем: нет никакой необходимости в воспроизведении того или иного *реального* хаотического окружения, вполне достаточно воспроизвести окружение *типичное*. Например, когда мы хотим узнать погоду на завтра, насколько *точная* информация нам в действительности нужна? Не сгодится ли *любое* правдоподобное описание?



## 3.11. Как обучаются роботы?

Учитывая вышесказанное, предлагаю остановиться на том, что на самом деле нас сейчас интересуют отнюдь не проблемы численного моделирования окружения. В принципе, возможностей поработать с окружением у нас будет предостаточно — *но только в том случае, если* не возникнет никаких трудностей с моделированием *внутренних* правил самой робототехнической системы. Поэтому перейдем к вопросу о том, как мы видим себе обучение нашего робота. Какие вообще процедуры обучения доступны вычислительному роботу? Возможно, ему будут предварительно заданы некие четкие правила вычислительного характера, как это обычно делается в нынешних системах на основе искусственных нейронных сетей (см. § 1.5). Такие системы подразумевают наличие некоторого четко определенного набора вычислительных правил, в соответствии с которыми усиливаются или ослабляются связи между составляющими сеть «нейронами», посредством чего достигается улучшение качества общего функционирования системы согласно критериям (искусственным или естественным), задаваемым внешним окружением. Еще один тип систем обучения образуют так называемые «генетические алгоритмы» — нечто вроде естественного отбора (или, если хотите, «выживания наиболее приспособленных») среди различных алгоритмических процедур, выполняемых на одной вычислительной машине; посредством такого отбора выявляется наиболее эффективный в управлении системой алгоритм.

Следует пояснить, что упомянутые правила (что характерно для восходящей организации вообще) несколько отличаются от стандартных нисходящих вычислительных алгоритмов, действующих в соответствии с известными процедурами для отыскания точных решений математических проблем. Восходящие правила лишь направляют систему к некоему общему улучшению качества ее функционирования. Впрочем, это не мешает им оставаться целиком и полностью алгоритмическими — в смысле воспроизводимости на универсальном компьютере (машине Тьюринга).

В дополнение к четким правилам такого рода, в совокупность средств, с помощью которых наша робототехническая система будет модифицировать свою работу, могут быть включены и некоторые *случайные* элементы. Возможно, эти случайные составляющие будут вноситься посредством каких-нибудь физических процессов — например, такого квантовомеханического процесса, как распад ядер радиоактивных атомов. На практике при конструировании искусственных вычислительных устройств имеет место тенденция к введению какой-либо вычислительной процедуры, результат вычисления в которой является случайным *по существу* (иначе такой результат называют *псевдослучайным*), хотя на деле он полностью определяется детерминистским характером самого вычисления (см. § 1.9). С описанным способом тесно связан другой, суть которого заключается в точном указании *момента времени*, в который производится вызов «случайной» величины, и введении затем этого момента времени в сложную вычислительную процедуру, которая и сама является, по существу, хаотической системой, вследствие чего малейшие изменения во времени дают эффективно непредсказуемые различия в результатах, а сами результаты становятся эффективно случайными. Хотя, строго говоря, наличие случайных компонентов и выводит рассматриваемые процедуры за рамки определения «операции машины Тьюринга», каких-то *существенных* изменений это за собой не влечет. В том, что касается функционирования нашего робота, случайным входным данным на практике оказываются эквивалентны псевдослучайные, а псевдослучайные входные данные *ничуть* не противоречат возможностям машины Тьюринга.

«Ну и что, что *на практике* случайные входные данные не отличаются от псевдослучайных? — заметит дотошный читатель. — *Принципиальная*-то разница между ними есть». На более раннем этапе нашего исследования (см., в частности, §§ 3.2–3.4) нас и в самом деле занимало то, чего математики могут достичь в принципе, вне зависимости от их практических возможностей. Более того, в определенных математических ситуациях проблему можно решить исключительно с помощью действительно случайных входных данных, никакие псевдослучайные заместители для этого не годятся. Подобные ситуации возникают, когда проблема подразумевает наличие некоего «состязательного» элемента, как часто бывает, например, в теории игр и криптографии. В некоторых видах «игр на двоих» оптимальная стратегия для каждого из игроков включает в себя, помимо прочего, и полностью случайную составляющую[9]. Любое сколько-нибудь последовательное пренебрежение одним из игроков необходимым для построения оптимальной стратегии элементом случайности позволяет другому игроку на протяжении достаточно длинной серии игр получить преимущество — по крайней мере, в принципе. Преимущество может быть достигнуто и в том случае, если противнику каким-то образом удалось составить достаточно достоверное представление о природе псевдослучайной (или иной) стратегии, используемой первым игроком вместо требуемой случайной. Аналогичным образом дело обстоит и в криптографии, где надежность кода напрямую зависит от того, насколько случайной является применяемая последовательность цифр. Если эта последовательность генерируется не истинно случайным образом, а посредством какого-либо псевдослучайного процесса, то, как и в случае с играми, этот процесс может в точности воспроизвести кто угодно, в том числе и потенциальный взломщик.



Поскольку случайность, как выясняется, представляет собой весьма ценное качество в таких состязательных ситуациях, то, на первый взгляд, можно предположить, что и в естественном отборе она должна играть не последнюю роль. Я даже уверен, что случайность и впрямь является во многих отношениях весьма важным фактором в процессе развития живых организмов. И все же, как мы убедимся несколько позднее в этой главе, одной лишь случайности оказывается недостаточно для того, чтобы вырваться из гёделевских сетей. И самые что ни на есть *подлинно* случайные элементы не помогут нашему роботу избежать ограничений, присущих вычислительным системам. Более того, у *псевдо*случайных процессов в этом смысле даже больше шансов, нежели у процессов чисто случайных (см. § 3.22).

Допустим на некоторое время, что наш робот и в самом деле является, по существу, *машиной Тьюринга* (хотя и с конечной емкостью запоминающего устройства). Строго говоря, учитывая, что робот непрерывно взаимодействует со своим окружением, а это окружение, как мы предполагаем, также допускает численное моделирование, было бы правильнее принять за единую машину Тьюринга робота *вместе* с окружением. Однако в целях удобства изложения я все же предлагаю рассматривать отдельно робота, как собственно машину Тьюринга, и отдельно окружение, как источник информации, поступающей на входную часть ленты машины. Вообще-то такую аналогию нельзя считать вполне приемлемой по одной формальной причине — машина Тьюринга есть устройство *фиксированное* и по определению неспособное изменять свою структуру «по мере накопления опыта». Можно, конечно, попытаться изобрести способ, посредством которого машина Тьюринга сможет-таки изменить свою структуру, — например, заставить машину работать безостановочно, модифицируя структуру в процессе работы, для чего непрерывно подавать на ее вход информацию от окружения. К нашему разочарованию, этот способ не сработает, поскольку *результат* работы машины Тьюринга можно узнать только после того, как машина достигнет внутренней команды STOP (см. § 2.1 и Приложение А, а также НРК, глава 2), после чего она не будет ничего считывать с входной части своей ленты до тех пор, пока мы не запустим ее снова. Когда же мы ее запустим, для продолжения работы ей придется возвратиться в исходное состояние, т. е. «обучиться» таким способом она ничему не сможет.

Впрочем, эту трудность можно обойти при помощи сложной технической модификации. Наша машина Тьюринга так и остается фиксированной, однако после каждого рабочего цикла, т. е. после достижения команды STOP, она дает на выходе два результата (формально кодируемые в виде одного-единственного числа). Первый результат определяет, каким в действительности будет ее последующее внешнее поведение, тогда как второй результат предназначен исключительно для *внутреннего* использования — в нем кодируется весь опыт, который машина получила от предыдущих контактов с окружением. В начале следующего цикла с входной части ее ленты *сначала* считывается «внутренняя» информация и *только после* нее все «внешние» данные, которыми машину снабжает окружение, включая и подробную



реакцию упомянутого окружения на ее предшествующее поведение. Таким образом, все результаты обучения оказываются записанными на, скажем так, *внутреннем* участке ленты, который машина в каждом рабочем цикле считывает заново (и который с каждым циклом становится все длиннее и длиннее).

## 3.12. Способен ли робот на «твердые математические убеждения»?

Воспользовавшись вышеописанным способом, мы и в самом деле можем представить себе в высшей степени обобщенного самообучающегося вычислительного «робота» в виде машины Тьюринга. Далее, предполагается, что наш робот способен судить об истинности математических утверждений, пользуясь при этом всеми способностями, потенциально присущими математикам-людям. И как же он будет это делать? Вряд ли нас обрадует необходимость кодировать каким-нибудь исключительно «нисходящим» способом все математические правила (все те, что входят в формальную систему $\mathbb{ZF}$, плюс все те, что туда не входят, о чем мы говорили выше), которые понадобятся роботу для того, чтобы иметь возможность непосредственно формировать собственные суждения подобно тому, как это делают люди, исходя из известных им правил, — поскольку, как мы могли убедиться, не существует ни одного сколько-нибудь приемлемого способа (за исключением, разумеется, «божественного вмешательства» — см. §§ 3.5, 3.6), посредством которого можно было бы реализовать такой неимоверно сложный и непознаваемо эффективный нисходящий алгоритм. Следует, очевидно, допустить, что какими бы внутренними «нисходящими» элементами ни обладал наш робот, они не являются жизненно важными для решения сложных математических проблем, а представляют собой всего лишь общие правила, обеспечивающие, предположительно, почву для формирования такого свойства как «понимание».

Выше (см. § 3.9) мы говорили о двух различных категориях входных данных, которые могут оказать существенное влияние на поведение нашего робота: *искусственных* и *естественных*. В качестве искусственного аспекта окружения мы рассматриваем учителя (одного или нескольких), который сообщает роботу о различных математических истинах и старается подтолкнуть его

к выработке каких-то внутренних критериев, с помощью которых робот мог бы самостоятельно отличать истинные утверждения от ложных. Учитель может информировать робота о совершенных тем ошибках или рассказывать ему о всевозможных математических понятиях и различных допустимых методах математического доказательства. Конкретные процедуры, применяемые в процессе обучения, учитель выбирает по мере необходимости из широкого диапазона возможных вариантов: «упражнение», «объяснение», «наставление» и даже, возможно, «порка». Что до естественных аспектов физического окружения, то они отвечают за «идеи», возникающие у робота в процессе наблюдения за поведением физических объектов; кроме того, окружение предоставляет роботу конкретные примеры воплощения различных математических понятий — например, понятия натуральных чисел: два апельсина, семь бананов, четыре яблока, один носок, ни одного ботинка и т. д., — а также хорошие приближения идеальных геометрических объектов (прямая, окружность) и некоторых бесконечных множеств (например, множество точек, заключенных внутри окружности).

Поскольку наш робот избежал-таки предварительного, полностью нисходящего программирования и, как мы предполагаем, формирует собственное понятие о математической истине с помощью всевозможных обучающих процедур, то нам следует позволить ему совершать в процессе обучения *ошибки* — с тем, чтобы он мог *учиться* и на своих ошибках. Первое время, по крайней мере, на эти ошибки ему будет указывать учитель. Кроме того, робот может самостоятельно обнаружить из наблюдений за окружением, что какие-то из его предыдущих, предположительно истинных математических суждений оказываются в действительности ошибочными, либо сомнительными и подлежащими повторной проверке. Возможно, он придет к такому выводу, основываясь исключительно на собственных соображениях о противоречивости этих своих суждений и т. д. Идея такова, что по мере накопления опыта робот будет делать все меньше и меньше ошибок. С течением времени учителя и физическое окружение будут становиться для робота все менее необходимыми — возможно, в конечном счете, окажутся и вовсе ненужными, — и при формировании своих математических суждений он будет все в большей степени опираться на собственную вычислительную мощь. Соответственно, можно предположить, что в дальнейшем



наш робот не ограничится теми математическими истинами, что он узнал от учителей или вывел из наблюдений за физическим окружением. Возможно, впоследствии он даже внесет какой-либо оригинальный вклад в математические исследования.

Для того чтобы оценить степень правдоподобия нарисованной нами картины, необходимо соотнести ее с теми вещами, что мы обсуждали ранее. Если мы хотим, чтобы наш робот и в самом деле обладал всеми способностями, пониманием и проницательностью математика-человека, ему потребуется какая-никакая концепция «неопровержимой математической истины». Его ранние попытки в формировании суждений, исправленные учителями или обесцененные наблюдением за физическим окружением, в эту категорию никоим образом не попадают. Они относятся к категории «догадок», а догадкам позволяется быть предварительными, пробными и даже ошибочными. Если предполагается, что наш робот должен вести себя как подлинный математик, то даже те ошибки, которые он будет порой совершать, должны быть исправимыми — причем, в принципе, исправимыми именно в соответствии с его собственными внутренними критериями «неопровержимой истинности».

Выше мы уже убедились, что концепцию «неопровержимой истины», которой руководствуется в своей деятельности математик-человек, нельзя сформировать посредством какого бы то ни было познаваемого (человеком) набора механических правил, в справедливости которых этот самый человек может быть целиком и полностью уверен. Если мы полагаем, что наш робот способен достичь уровня математических способностей, достижимого, *в принципе*, для любого человеческого существа (а то и превзойти этот уровень), то в этом случае *его* (робота) концепция неопровержимой математической истины также должна представлять собой нечто такое, что невозможно воспроизвести посредством набора механических правил, которые можно полагать обоснованными, — т. е. правил, которые может полагать обоснованными математик-человек или, коли уж на то пошло, математик-робот.

В связи с этими соображениями возникает один весьма важный вопрос: *чьи* же концепции, восприятие, неопровержимые убеждения следует считать значимыми — наши или роботов? Можно ли полагать, что робот *действительно* обладает убеждениями или способен что-либо осознавать? Если читатель

придерживается точки зрения $\mathscr{B}$, то он, возможно, сочтет такой вопрос несколько неуместным, поскольку сами понятия «осознания» или «убеждения» относятся к описанию процесса *мышления* и поэтому никоим образом неприменимы к целиком компьютерному роботу. Однако в рамках настоящего рассуждения нет необходимости в том, чтобы наш гипотетический робот и в самом деле обладал какими-то подлинными ментальными качествами, коль скоро мы допускаем, что он способен *внешне* вести себя в точности подобно математику-человеку — в полном соответствии с самыми строгими формулировками как $\mathscr{B}$, так и $\mathscr{A}$. Нам не нужно, чтобы робот *действительно* понимал, осознавал или верил; достаточно того, что внешне он проявляет себя в точности так, будто он этими ментальными качествами в полной мере обладает. Подробнее об этом мы поговорим в § 3.17.

Точка зрения $\mathscr{B}$ не отличается принципиально от $\mathscr{A}$ в том, что касается ограничений, налагаемых на возможную манеру поведения робота, однако сторонники $\mathscr{B}$, скорее всего, питают несколько меньшие *надежды* в отношении тех высот, которых на деле может достичь робот, или вероятности создания вычислительной системы, которую можно было бы полагать способной на эффективное моделирование деятельности мозга человека, оценивающего обоснованность того или иного математического рассуждения. Подобное *человеческое* восприятие предполагает все же некоторое понимание *смысла* затронутых математических концепций. Согласно точке зрения $\mathscr{A}$, во всем этом нет ничего, выходящего за рамки некоторого свойства вычисления, связанного с понятием «смысла», тогда как $\mathscr{B}$ рассматривает смысл в качестве семантического аспекта мышления и не допускает возможности его описания в чисто вычислительных терминах. В этом мы согласны с точкой зрения $\mathscr{B}$ и отнюдь не ожидаем от нашего робота способности действительно ощущать тонкие семантические различия. Таким образом, сторонники $\mathscr{B}$, возможно, менее (нежели сторонники $\mathscr{A}$) склонны предполагать, что какой бы то ни было робот, сконструированный в соответствии с обсуждаемыми здесь принципами, окажется когда-либо способен на демонстрацию тех внешних проявлений человеческого понимания, какие свойственны математикам-людям. Полагаю, отсюда можно сделать вывод (не такой, собственно, и неожиданный), что сторонников $\mathscr{B}$ будет существенно легче обратить в приверженцев $\mathscr{C}$, чем сторонников $\mathscr{A}$; впрочем, для нашего дальнейшего



исследования разница между $\mathscr{A}$ и $\mathscr{B}$ существенного значения не имеет.

В качестве заключения отметим, что, хотя истинность математических утверждений нашего робота, получаемых посредством преимущественно восходящей системы вычислительных процедур, носит заведомо предварительный и предположительный характер, следует допустить, что роботу действительно присущ некоторый достаточно «прочный» уровень *неопровержимой* математической «убежденности», вследствие чего некоторые из его утверждений (которым он будет присваивать некий особый статус — обозначаемый, скажем, знаком ☆) нужно считать неопровержимо истинными — согласно *собственным* критериям робота. О допустимости ошибочного присвоения роботом статуса ☆ — пусть роботом же и исправимом — мы поговорим в § 3.19. А до той поры будем полагать, что всякое ☆-утверждение робота следует рассматривать как безошибочное.

## 3.13. Механизмы математического поведения робота

Рассмотрим различные механизмы, лежащие в основе процедур, управляющих поведением робота в процессе получения им ☆-утверждений. Некоторые из этих процедур являются по отношению к роботу *внутренними* — нисходящие внутренние ограничители, встроенные в модель функционирования робота, а также те или иные заранее определенные восходящие процедуры, посредством которых робот улучшает качество своей работы (с тем, чтобы постепенно достичь ☆-уровня). Разумеется, мы полагаем, что все эти процедуры в принципе познаваемы человеком (хотя окончательный результат совокупного действия всех этих разнообразных факторов вполне может оказаться за пределами вычислительных способностей математика-человека). В самом деле, если мы допускаем, что человеческие существа в один прекрасный день сконструируют робота, наделенного подлинным математическим талантом, то следует непременно допустить и то, что человек способен понять внутренние принципы, в соответствии с которыми будет построен этот робот, иначе любое подобное начинание обречено на провал.

Безусловно, мы отдаем себе отчет в том, что создание такого робота вполне может оказаться многоступенчатым процессом:

иначе говоря, возможно, что наш робот-математик будет целиком и полностью построен какими-либо роботами «низшего порядка» (которые сами не способны на подлинно математическую деятельность), а эти роботы, в свою очередь, построены другими роботами еще более низкого порядка. Однако запущена в производство вся эта иерархическая цепочка будет все равно человеком, и исходные правила ее построения (по всей видимости, некая комбинация нисходящих и восходящих процедур) будут в любом случае доступны человеческому пониманию.

Существенно важными для процесса развития робота являются и всевозможные *внешние* факторы, привносимые окружением. Внешний мир и в самом деле может обеспечить нашего робота весьма значительным объемом вводимых данных, поступающих как от учителей-людей (или роботов), так и из наблюдений за естественным физическим окружением. Что до естественных внешних факторов, привносимых «безлюдным» окружением, то «непознаваемыми» их, как правило, не считают. Эти факторы могут быть очень сложными, часто они взаимодействуют между собой, и все же эффективное «виртуально-реальное» моделирование существенных аспектов нашего окружения уже вполне осуществимо (см. § 1.20). По-видимому, ничто не мешает модифицировать эти модели таким образом, чтобы робот с их помощью получал все, что ему нужно для развития в смысле внешних естественных факторов, — не будем забывать, что вполне достаточно смоделировать *типичное* окружение, воспроизводить какое-то реально существующее необходимости нет (см. §§ 1.7, 1.9).

Вмешательство в процесс людей (или роботов) — т. е. внешних, «искусственных» факторов — может происходить на различных этапах, однако это никоим образом не влияет на существенную познаваемость механизмов этого вмешательства, при условии, разумеется, что мы допускаем возможность каким-то познаваемым образом «механизировать» вмешательство человека. Справедливо ли такое допущение? Думаю, вполне естественно (по крайней мере, для сторонника точки зрения $\mathscr{A}$ или $\mathscr{B}$) предположить, что любое человеческое вмешательство в процесс развития робота и в самом деле можно заменить какими-либо целиком и полностью вычислительными процедурами. Мы же не требуем, чтобы в этом вмешательстве непременно присутствовало что-либо непостижимо мистическое — скажем, некая неопределимая «сущность», какую учитель-человек должен передать своему ученику-роботу в процессе обучения. Мы полагаем, что при обучении роботу необходимо получать всего лишь те или иные фундаментальные сведения, а передачу ему этих сведений проще всего поручить именно человеку. Весьма вероятно, что, как и в случае с учениками-людьми, наиболее эффективной будет передача информации в интерактивной форме, когда поведение учителя зависит от реакции ученика. Однако и это обстоятельство, само по себе, отнюдь не исключает возможности эффективно вычислительного поведения учителя. В конце концов, все наши рассуждения в настоящей главе представляют собой одно сплошное *reductio ad absurdum*, в рамках которого мы допускаем, что в поведении человеческих существ вообще нет ничего существенно невычислимого. А тем, кто уже и так придерживается точек зрения $\mathscr{C}$ или $\mathscr{D}$ (последние, несомненно, склонны скорее поверить в возможность существования упомянутой выше невычислимой «сущности», передаваемой роботу в силу одного лишь человеческого происхождения учителя), наши доказательства в любом случае совершенно не нужны.



Если рассматривать все эти механизмы (т. е. внутренние вычислительные процедуры и данные, поступающие от интерактивного внешнего окружения) в совокупности, то создается впечатление, что нет каких-либо разумных причин полагать их принципиально непознаваемыми, — даже если кто-то и настаивает на том, что на практике в точности просчитать результирующие проявления внешних из упомянутых механизмов не в силах человеческих (и даже не в силах любого из существующих или предвидимых в обозримом будущем компьютеров). К вопросу о познаваемости вычислительных механизмов мы еще вернемся, причем довольно скоро (в конце § 3.15). А пока допустим, что все эти механизмы действительно познаваемы, и обозначим набор таких механизмов буквой **M**. Возможно ли, что некоторые из полученных с помощью этих механизмов утверждений ☆-уровня окажутся, тем не менее, *не*познаваемыми для человека? Обоснованно ли такое предположение? Вообще говоря, нет — при условии, что в данном контексте мы продолжаем интерпретировать понятие «познаваемости» в том же *принципиальном* смысле, который мы применяли в отношении случаев **I** и **II** и который был исчерпывающе определен в начале § 3.5. Тот факт, что нечто (например, формулировка некоего ☆-утверждения) может оказать-

ся за пределами *невооруженных* вычислительных способностей человеческого существа, к данному случаю отношения не имеет. Ничуть не возбраняется и «вооружить» человека теми или иными средствами содействия мыслительным процессам — например, карандашом и бумагой, карманным калькулятором либо универсальным компьютером в комплекте с программным обеспечением нисходящего типа. Даже если добавить к уже имеющимся вычислительным процедурам какие-либо восходящие компоненты, то мы не получим ничего такого, чего не могли бы *в принципе* получить раньше — при условии, разумеется, что лежащие в основе этих восходящих процедур фундаментальные *механизмы* доступны человеческому пониманию. С другой стороны, вопрос о «познаваемости» самих механизмов **M** следует рассматривать уже в «практическом» смысле — в полном соответствии с принятой в § 3.5 терминологией. Таким образом, на данный момент мы полагаем, что механизмы **M** являются действительно познаваемыми *практически*.

Обладая знанием механизмов **M**, мы можем использовать их при создании фундамента для построения *формальной системы* $\mathbb{Q}(\mathbf{M})$, при этом *теоремами* такой системы станут следующие положения: (i) ☆-утверждения, непосредственно следующие из применения упомянутых механизмов, и (ii) любые положения, выводимые из этих ☆-утверждений с применением правил элементарной логики. Под «элементарной логикой» здесь могут пониматься, скажем, правила *исчисления предикатов* (описанные в § 2.9) или какая-либо иная столь же прямая и четко определенная неопровержимая система аналогичных логических правил (вычислительных). Мы вполне способны построить формальную систему $\mathbb{Q}(\mathbf{M})$ в силу того простого факта, что процедура $Q(\mathbf{M})$, посредством которой из набора механизмов **M** получаются, одно за другим, необходимые ☆-утверждения, является процедурой *вычислительной* (пусть на практике и весьма громоздкой). Отметим, что определяемая таким образом процедура $Q(\mathbf{M})$ будет генерировать утверждения группы (i), однако вовсе не обязательно все положения группы (ii) (поскольку можно допустить, что нашему роботу, по всей вероятности, попросту надоест тупо выводить все логические следствия из вырабатываемых им ☆-теорем). Таким образом, процедура $Q(\mathbf{M})$ не эквивалентна в точности формальной системе $\mathbb{Q}(\mathbf{M})$, однако различие между ними не существенно. К тому же ничто не мешает нам при желании получить из процедуры $Q(\mathbf{M})$ другую процедуру — такую, например, которая *будет* эквивалентна $\mathbb{Q}(\mathbf{M})$.



Далее, для интерпретации формальной системы $\mathbb{Q}(\mathbf{M})$ необходимо каким-то образом устроить так, чтобы на всем протяжении развития робота статус ☆ всегда и непременно *означал*, что удостоенное его утверждение действительно следует полагать неопровержимо доказанным. В отсутствие поступающих от учителя-человека (неважно, в какой форме) внешних данных мы не можем быть уверенными в том, что робот не выработает самостоятельно некий отличный от нашего язык, в котором символ ☆ будет иметь совершенно иное значение (либо вовсе окажется бессмысленным). Для того чтобы определение формальной системы $\mathbb{Q}(\mathbf{M})$ на языке робота согласовывалось с нашим ее определением, необходимо в процессе обучения робота (например, учителем-человеком) проследить за тем, чтобы присваиваемое символу ☆ значение в точности соответствовало тому значению, какое в него вкладываем мы. Необходимо также проследить и за тем, чтобы система обозначений, которой робот фактически пользуется при формулировке своих, скажем, $\Pi_1$-высказываний, в точности совпадала с аналогичной системой, имеющей хождение у нас (или допускала какое-либо явное преобразование в нашу систему). Если допустить, что механизмы **M** познаваемы человеком, то из вышесказанного следует, что аксиомы и правила действия формальной системы $\mathbb{Q}(\mathbf{M})$ также должны быть познаваемыми. Более того, всякую теорему, выводимую в рамках системы $\mathbb{Q}(\mathbf{M})$, следует, *в принципе*, полагать познаваемой человеком (в том смысле, что мы в состоянии понять ее описание, а не определить в обязательном порядке ее неопровержимую истинность), даже если вычислительные процедуры, необходимые для получения большей части таких теорем, окажутся далеко за пределами невооруженных вычислительных способностей человека.

## 3.14. Фундаментальное противоречие

Предшествующая дискуссия в сущности показывает, что «непознаваемый и неосознаваемый алгоритм $F$», который, согласно допущению **III**, лежит в основе восприятия математической истины, вполне возможно свести к алгоритму осознанно познаваемому — при условии, что нам, следуя заветам адептов ИИ,

удастся запустить некую систему процедур, которые в конечном счете приведут к созданию робота, способного на математические рассуждения на человеческом (а то и выше) уровне. Непознаваемый алгоритм $F$ заменяется при этом вполне познаваемой формальной системой $\mathbb{Q}(\mathbf{M})$.

Прежде чем мы приступим к подробному рассмотрению этого аргумента, необходимо обратить внимание на один существенный момент, который мы до сих пор незаслуженно игнорировали — речь идет о возможности привнесения на разных этапах процесса развития робота неких *случайных элементов* взамен раз и навсегда фиксированных механизмов. В свое время нам еще предстоит обратиться к этому вопросу, пока же я буду полагать, что каждый такой случайный элемент следует рассматривать как результат выполнения какого-либо *псевдо*случайного (хаотического) вычисления. Как было показано ранее (§§ 1.9, 3.11), таких псевдослучайных компонентов на практике оказывается вполне достаточно. К случайным элементам в «образовании» робота мы еще вернемся в § 3.18, где более подробно поговорим о подлинной случайности в применении к нашему случаю, а пока, говоря о «наборе механизмов $\mathbf{M}$», я буду предполагать, что все эти механизмы действительно являются целиком и полностью вычислительными и свободными от какой бы то ни было реальной неопределенности.

Суть противоречия заключается в том, что на месте алгоритма $F$, фигурировавшего в наших предыдущих рассуждениях (например, того алгоритма, о котором мы говорили в § 3.2 в связи с допущением **I**), с неизбежностью оказывается формальная система $\mathbb{Q}(\mathbf{M})$. Вследствие чего случай **III** эффективно сводится к случаю **I** и тем самым не менее эффективно из рассмотрения исключается. Выступая в рамках данного доказательства в роли сторонников точек зрения $\mathscr{A}$ и $\mathscr{B}$, мы предполагаем, что наш робот *в принципе* способен (с помощью обучающих процедур той же природы, что установили для него мы) достичь в конечном счете любых математических результатов, каких в состоянии достичь человек. Мы должны также допустить, что робот *способен* достичь и таких результатов, какие человеку в принципе *не по силам*. Так или иначе, нашему роботу предстоит обзавестись способностью к пониманию мощи аргументации Гёделя (или, по крайней мере, способностью *сымитировать* такое понимание — согласно $\mathscr{B}$). Иначе говоря, относительно любой заданной



(достаточно обширной) формальной системы $\mathbb{H}$ робот должен оказаться в силах неопровержимо установить тот факт, что из обоснованности системы $\mathbb{H}$ следует истинность его гёделевского[5] утверждения $G(\mathbb{H})$, а также то, что утверждение $G(\mathbb{H})$ не является теоремой системы $\mathbb{H}$. В частности, робот сможет установить, что из обоснованности системы $\mathbb{Q}(\mathbf{M})$ неопровержимо следует истинность утверждения $G(\mathbb{Q}(\mathbf{M}))$; эта же обоснованность предполагает, что утверждение $G(\mathbb{Q}(\mathbf{M}))$ не является теоремой системы $\mathbb{Q}(\mathbf{M})$.

С помощью в точности тех же рассуждений, какими мы воспользовались в § 3.2 применительно к человеческому математическому пониманию, непосредственно из вышеизложенных соображений выводится, что робот никоим образом не способен твердо поверить в то, что совокупность его собственных — и, на его взгляд, неопровержимых — математических убеждений *действительно* эквивалентна некоей формальной системе $\mathbb{Q}(\mathbf{M})$. И это несмотря на тот факт, что *мы* (выступая в роли соответствующих экспертов по проблемам ИИ) прекрасно осведомлены о том, что в основе системы математических убеждений робота лежит не что-нибудь, а именно набор механизмов $\mathbf{M}$, что автоматически означает, что система неопровержимых убеждений робота *является* полным эквивалентом системы $\mathbb{Q}(\mathbf{M})$. Если бы робот вдруг твердо поверил в то, что все его убеждения укладываются в рамки системы $\mathbb{Q}(\mathbf{M})$, то тогда ему пришлось бы поверить и в обоснованность этой самой системы $\mathbb{Q}(\mathbf{M})$. Соответственно, ему также пришлось бы одновременно поверить и в истинность утверждения $G(\mathbb{Q}(\mathbf{M}))$, и в то, что упомянутое утверждение в его систему убеждений не входит — неразрешимое противоречие! Иначе говоря, робот никак не может знать о том, что он сконструирован в соответствии с тем или иным набором механизмов $\mathbf{M}$. А поскольку об *этой* особенности его конструкции знаем — или по крайней мере, в состоянии узнать — мы с вами, то получается, что нам доступны такие математические истины (например, утверждение $G(\mathbb{Q}(\mathbf{M}))$), которые роботу оказываются не по силам, хотя изначально предполагалось, что способности робота будут равны способностям человека (или даже превысят их).

[5] В ранних изданиях этой книги вместо обозначения $G(\mathbb{F})$ в оставшейся части главы 3 использовалось обозначение $\Omega(\mathbb{F})$. Однако $G(\mathbb{F})$, на мой взгляд, представляется в данном случае более уместным (см. также § 2.8 и с. 160).

## 3.15. Способы устранения фундаментального противоречия

Приведенное выше рассуждение можно рассматривать двояко — с точки зрения создавших робота людей либо с точки зрения самого робота. С человеческой точки зрения существует некоторая неопределенная вероятность того, что математику-человеку претензии робота на обладание неопровержимой истиной покажутся неубедительными, разве что упомянутый математик-человек примет во внимание какие-то отдельные конкретные *аргументы* из тех, что использует робот. Возможно, не все теоремы системы $\mathbb{Q}(\mathbf{M})$ человек сочтет неопровержимо истинными, кроме того, как нам помнится, интеллектуальные способности робота могут существенно *превышать* таковые же способности человека. Таким образом, можно утверждать, что одно лишь знание о том, что робот сконструирован в соответствии с неким набором механизмов $\mathbf{M}$, не следует рассматривать в качестве неопровержимо убедительной (для человека) математической демонстрации. Соответственно, мы должны пересмотреть все вышеприведенное рассуждение — на этот раз с точки зрения *робота*. Какие огрехи в нашем обосновании в состоянии заметить (и использовать) робот?

По-видимому, наш робот располагает всего лишь четырьмя основными возможностями для нейтрализации фундаментального противоречия — при условии, конечно, что сам робот осведомлен о том, что он является в некотором роде вычислительной машиной.

(a) *Возможно*, что робот, принимая в целом утверждение о том, что в основе его конструкции лежит некий набор механизмов $\mathbf{M}$, тем не менее, неизбежно остается неспособен *безоговорочно* поверить в этот факт.

(b) Возможно, что робот, будучи безоговорочно убежден в истинности каждого отдельного ☆-утверждения в тот момент, когда он его формулирует, все же сомневается в достоверности *полной* системы своих ☆-утверждений — соответственно, робот может не верить в то, что формальная система $\mathbb{Q}(\mathbf{M})$ и *в самом деле* лежит в основе всей его системы убеждений в отношении $\Pi_1$-высказываний.



(c) Возможно, что подлинный набор механизмов $\mathbf{M}$ существенно зависит от *случайных* элементов и не может быть адекватно описан через посредство неких известных результатов псевдослучайных вычислений, подаваемых на входное устройство робота.

(d) Возможно, что подлинный набор механизмов $\mathbf{M}$ в действительности *непознаваем*.

В последующих девяти разделах представлен ряд веских аргументов, убедительно демонстрирующих, что первые три лазейки ((a), (b) и (c)) оказываются для робота, задавшегося целью обойти фундаментальное противоречие, совершенно бесполезными. Соответственно, робот (а вместе с ним и мы — если мы, конечно, продолжаем настаивать на том, что математическое понимание можно свести к вычислению) начинает всерьез подумывать о не очень привлекательной возможности (d). Уверен, что непривлекательной возможность (d) нахожу не я один — думаю, в этом со мной согласятся и те читатели, которым не безразлична судьба идеи искусственного интеллекта. Ее, пожалуй, приемлемо рассматривать лишь в качестве возможной мировоззренческой позиции, укладывающейся, по сути своей, в рамки той самой комбинации точек зрения $\mathscr{A}$ и $\mathscr{D}$, о которой мы говорили в конце § 1.3 и согласно которой для внедрения непознаваемого алгоритма в «мозг» каждого из наших роботов требуется, ни много ни мало, *божественное вмешательство* (от «первого в мире программиста»). В любом случае, вердикт «непознаваемо», вынесенный в отношении тех самых механизмов, которые, в конечном счете, ответственны за наличие у нас какого ни на есть разума, вряд ли обрадует тех, кто намерен, вообще говоря, *построить* робота, наделенного подлинным искусственным интеллектом. Не особенно обрадует он и тех из нас, кто все еще надеется понять, принципиально и не выходя за рамки строго научного подхода, каким образом в действительности возникло у человека такое свойство, как интеллект, объяснить его происхождение посредством четко формулируемых научных законов — законов физики, химии, биологии, законов естественного отбора, в конце концов, — пусть даже и не имея в виду воспроизвести этот самый интеллект в каком бы то ни было робототехническом устройстве. Лично я полагаю, что подобный пессимистический вердикт не имеет под собой никаких оснований — по той хотя бы простой причине, что

«научная постижимость» имеет весьма мало общего с «вычислимостью». Законы, лежащие в основе мыслительных процессов не являются непостижимыми, они всего лишь *невычислимы*. На эту тему мы еще поговорим во второй части книги.

## 3.16. Необходимо ли роботу верить в механизмы M?

Вообразим, что у нас имеется робот, снабженный некоторым возможным набором механизмов **M**, — каковой набор может оказаться тем самым, на основе которого и построен наш робот, но это не обязательно. Я попробую убедить читателя в том, что робот будет вынужден отвергнуть возможность того, что его математическое понимание опирается на набор механизмов **M**, — *независимо* от того, как обстоит дело в действительности. При этом мы на время допускаем, что робот по тем или иным причинам уже отбросил варианты (b), (c) и (d), и приходим к выводу (несколько даже неожиданному), что сам по себе вариант (a) избежать парадокса не позволяет.

Рассуждать мы будем следующим образом. Обозначим через $\mathscr{M}$ гипотезу

> «В основе математического понимания робота лежит набор механизмов **M**»

и рассмотрим утверждение вида

> «Такое-то $\Pi_1$-высказывание является следствием из $\mathscr{M}$».

Такое утверждение (в том случае, когда робот твердо верит в его истинность) я буду называть $\bigstar_{\mathscr{M}}$-утверждением. Иначе говоря, под $\bigstar_{\mathscr{M}}$-утверждениями не обязательно понимаются те $\Pi_1$-высказывания, в истинность которых как таковых неопровержимо верит робот, но те $\Pi_1$-высказывания, которые робот полагает неопровержимо выводимыми из гипотезы $\mathscr{M}$. Изначально от робота не требуется обладание какими бы то ни было взглядами относительно возможности того, что в основе его конструкции *действительно* лежит набор механизмов **M**. Он может даже поначалу счесть такое предположение абсолютно невероятным, но, тем не менее, ничто не мешает ему рассмотреть (в подлинно



научной традиции) возможные следствия из *гипотезы* о таком вот его происхождении.

Существуют ли $\Pi_1$-высказывания, которые робот должен полагать неопровержимыми следствиями из гипотезы $\mathscr{M}$ и которые при этом не являются самыми обыкновенными $\bigstar$-утверждениями, вовсе не требующими привлечения этой гипотезы? Разумеется, существуют. Как было отмечено в конце §3.14, истинность $\Pi_1$-высказывания $G(\mathbb{Q}(\mathbf{M}))$ следует из обоснованности формальной системы $\mathbb{Q}(\mathbf{M})$, отсюда же следует и тот факт, что утверждение $G(\mathbb{Q}(\mathbf{M}))$ не является теоремой системы $\mathbb{Q}(\mathbf{M})$. Более того, в этом робот будет совершенно безоговорочно убежден. Если допустить, что робот вполне согласен с тем, что все его неопровержимые убеждения укладывались бы в рамки системы $\mathbb{Q}(\mathbf{M})$, *будь* он действительно сконструирован в соответствии с набором механизмов **M**, — т. е. что возможность (b)[6] он из рассмотрения исключает, — то получается, что наш робот и в самом деле должен твердо верить в то, что обоснованность системы $\mathbb{Q}(\mathbf{M})$ является следствием гипотезы $\mathscr{M}$. Таким образом, робот оказывается безоговорочно убежден как в том, что $\Pi_1$-высказывание $G(\mathbb{Q}(\mathbf{M}))$ следует из гипотезы $\mathscr{M}$, так и в том, что (согласно $\mathscr{M}$) он не способен непосредственно постичь его неопровержимую истинность без привлечения $\mathscr{M}$ (поскольку формальной системе $\mathbb{Q}(\mathbf{M})$ оно не принадлежит). Соответственно, утверждение $G(\mathbb{Q}(\mathbf{M}))$ является $\bigstar_{\mathscr{M}}$-утверждением, но не $\bigstar$-утверждением.

Предположим, что формальная система $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ построена в точности так же, как и система $\mathbb{Q}(\mathbf{M})$, с той лишь разницей, что роль, которую при построении системы $\mathbb{Q}(\mathbf{M})$ исполняли $\bigstar$-утверждения, сейчас берут на себя $\bigstar_{\mathscr{M}}$-утверждения. Иначе говоря, теоремами системы $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ являются либо (i) сами $\bigstar_{\mathscr{M}}$-утверждения, либо (ii) положения, выводимые из этих $\bigstar_{\mathscr{M}}$-утверждений с применением правил элементарной логики (см. §3.13). Точно так же, как робот на основании гипотезы $\mathscr{M}$ согласен с тем, что формальная система $\mathbb{Q}(\mathbf{M})$ охватывает все его неопровержимые убеждения относительно истинности $\Pi_1$-высказываний, он будет согласен и с тем, что формальная

[6] Само собой разумеется, что вариант (d) мы в данном случае даже не рассматриваем, так как набор механизмов **M** был роботу в явном виде *предъявлен*, кроме того, мы на время допускаем, что механизмы **M** не включают в себя никаких случайных элементов, вследствие чего вариант (c) также отпадает.

система $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ охватывает все его неопровержимые убеждения относительно истинности $\Pi_1$-высказываний, обусловленных гипотезой $\mathscr{M}$.

Далее предложим роботу рассмотреть гёделевское $\Pi_1$-высказывание $G(\mathbb{Q}_{\mathscr{M}}(\mathbf{M}))$. Робот, несомненно, проникнется неопровержимым убеждением в том, что это $\Pi_1$-высказывание является следствием из обоснованности системы $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$. Он также вполне безоговорочно поверит в то, что обоснованность системы $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ является следствием гипотезы $\mathscr{M}$, поскольку он согласен с тем, что система $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ действительно содержит в себе все, в чем робот неопровержимо убежден в отношении своей способности выводить $\Pi_1$-высказывания, основываясь на гипотезе $\mathscr{M}$. (Он будет рассуждать следующим образом: «Если я принимаю гипотезу $\mathscr{M}$, то я тем самым принимаю и все $\Pi_1$-высказывания, которые порождают систему $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$. Таким образом, я должен согласиться с тем, что система $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ является обоснованной на основании гипотезы $\mathscr{M}$. Следовательно, на основании все той же гипотезы, я должен признать и то, что утверждение $G(\mathbb{Q}_{\mathscr{M}}(\mathbf{M}))$ истинно».)

Однако, поверив (безоговорочно) в то, что гёделевское $\Pi_1$-высказывание $G(\mathbb{Q}_{\mathscr{M}}(\mathbf{M}))$ является следствием гипотезы $\mathscr{M}$, робот вынужден будет поверить и в то, что утверждение $G(\mathbb{Q}_{\mathscr{M}}(\mathbf{M}))$ является теоремой формальной системы $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$. А в это он сможет поверить только в том случае, если он полагает систему $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ *необоснованной*, — что решительно противоречит принятию им гипотезы $\mathscr{M}$.

В некоторых из вышеприведенных рассуждений неявно допускалось, что неопровержимая убежденность робота является *действительно* обоснованной, — хотя необходимо лишь, чтобы сам робот просто верил в обоснованность своей системы убеждений. Впрочем, мы изначально предполагаем, что наш робот обладает математическим пониманием, по крайней мере, на человеческом уровне, а человеческое математическое понимание, как было показано в § 3.4, принципиально является обоснованным.

Возможно, кто-то усмотрит в формулировке допущения $\mathscr{M}$, равно как и в определении $\bigstar_{\mathscr{M}}$-утверждения, некоторую неоднозначность. Смею вас уверить, что подобное утверждение, будучи $\Pi_1$-высказыванием, представляет собой в высшей степени определенное математическое утверждение. Можно предположить, что большинство $\bigstar_{\mathscr{M}}$-утверждений робота окажутся



в действительности самыми обыкновенными $\bigstar$-утверждениями, поскольку маловероятно, что робот при каких угодно обстоятельствах сочтет целесообразным прибегать в своих рассуждениях к самóй гипотезе $\mathscr{M}$. Исключением может стать утверждение $G(\mathbb{Q}(\mathbf{M}))$, о котором говорилось выше, так как в данном случае формальная система $\mathbb{Q}(\mathbf{M})$ выступает, с точки зрения робота, в роли гёделевской гипотетической «машины для доказательства теорем» (см. §§ 3.1 и 3.3). Вооружившись гипотезой $\mathscr{M}$, робот получает доступ к своей собственной «машине для доказательства теорем», и, хотя он не может быть (да и, скорее всего, не будет) безоговорочно убежден в обоснованности своей «машины», робот способен предположить, что она *может* оказаться обоснованной, и попытаться вывести следствия уже из этого предположения.

На этом этапе робот еще не добирается до парадокса — так же, как не добрался до него и Гёдель в своих рассуждениях о человеческом интеллекте (см. цитату в § 3.1). Однако, поскольку роботу доступен для исследования набор гипотетических *механизмов* $\mathbf{M}$, а не просто отдельная формальная система $\mathbb{Q}(\mathbf{M})$, он может повторить свое рассуждение и перейти от системы $\mathbb{Q}(\mathbf{M})$ к системе $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$, обоснованность которой он по-прежнему полагает простым следствием из гипотезы $\mathscr{M}$. Именно *это* и приводит его в конечном итоге к противоречию (чего мы, собственно, и добивались). (См. также § 3.24, где мы продолжим рассмотрение системы $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ и ее кажущейся связи с «парадоксальными рассуждениями».)

Вывод: ни одно обладающее сознанием и имеющее понятие о математике существо — иначе говоря, ни одно существо со способностью к подлинному математическому пониманию — не может функционировать в соответствии с каким бы то ни было набором постижимых им механизмов, вне зависимости от того, *знает* ли оно в действительности о том, что именно *эти* механизмы, предположительно, направляют его на его пути к неопровержимой математической истине. (Вспомним и о том, что «неопровержимой математической истиной» это существо полагает всего лишь то, что оно способно установить математическими методами, — т. е. с помощью «математического доказательства», причем совсем необязательно «формального».)

Если конкретнее, то на основании предшествующих рассуждений мы склонны заключить, что не существует такого пости-

жимого роботом и не содержащего подлинно случайных компонентов набора вычислительных механизмов, какой робот мог бы принять (даже в качестве *возможности*) как основу своей системы математических убеждений, — *при условии*, что робот готов согласиться с тем, что специфическая процедура, предложенная мною для построения формальной системы $\mathbb{Q}(\mathbf{M})$ на основе механизмов $\mathbf{M}$, *и в самом деле* охватывает всю совокупность $\Pi_1$-высказываний, в истинность которых он неопровержимо верит, а также, соответственно, с тем, что формальная система $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ охватывает всю совокупность $\Pi_1$-высказываний, которые, как он неопровержимо верит, следуют из гипотезы $\mathscr{M}$. Кроме того, если мы хотим, чтобы робот смог построить собственную потенциально непротиворечивую систему математических убеждений, следует ввести в набор механизмов $\mathbf{M}$ какие-либо подлинно случайные составляющие.

Эти последние оговорки мы рассмотрим в последующих разделах (§§ 3.17–3.22). Вопрос о введении в набор механизмов $\mathbf{M}$ возможных случайных элементов (вариант (c)) представляется удобным обсудить в рамках общего рассмотрения варианта (b). А для того чтобы рассмотреть вариант (b) с должной тщательностью, нам следует прежде в полной мере прояснить для себя вопрос об «убежденности» робота, который мы уже мимоходом затрагивали в конце § 3.12.

## 3.17. Робот ошибается и робот «имеет в виду»?

Важнейший вопрос из тех, с какими нам предстоит разобраться на данном этапе, звучит так: готов ли робот безоговорочно согласиться с тем, что — при условии его построения в соответствии с некоторым набором механизмов $\mathbf{M}$ — формальная система $\mathbb{Q}(\mathbf{M})$ корректным образом включает в себя всю систему его математических убеждений в отношении $\Pi_1$-высказываний (равно как и с соответствующим предположением для системы $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$)? Такое согласие подразумевает, прежде всего, что робот верит в *обоснованность* системы $\mathbb{Q}(\mathbf{M})$, — т. е. в то, что все $\Pi_1$-высказывания, являющиеся ☆-утверждениями, действительно *истинны*. Наши рассуждения требуют также, чтобы *всякое* $\Pi_1$-высказывание, в истинность которого робот в состоянии безоговорочно поверить, являлось непременно теоремой системы $\mathbb{Q}(\mathbf{M})$ (т. е. чтобы в рамках системы $\mathbb{Q}(\mathbf{M})$ робот мог



бы определить «машину для доказательства теорем», аналогичную той, возможность создания которой в случае математиков-людей допускал Гёдель, см. §§ 3.1, 3.3). Вообще говоря, существенно *не* то, чтобы система $\mathbb{Q}(\mathbf{M})$ действительно играла такую универсальную роль в отношении потенциальных способностей робота, связанных с $\Pi_1$-высказываниями, а лишь то, чтобы она была достаточно обширна для того, чтобы допускать применение гёделевского доказательства к самой себе (и, соответственно, к системе $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$). Позднее мы увидим, что необходимость в таком применении возникает лишь в случае некоторых конечных систем $\Pi_1$-высказываний.

Таким образом, мы — как, собственно, и робот — должны учитывать возможность того, что некоторые из ☆-утверждений робота окажутся в действительности ошибочными, и то, что робот может самостоятельно обнаружить и исправить эти ошибки согласно собственным внутренним критериям, сути дела не меняет. А суть дела заключается в том, что поведение робота в этом случае становится как нельзя более похоже на поведение математика-человека. Человеку ничего не стоит оказаться в ситуации, когда он (или она) полагает, что истинность (или ложность) того или иного $\Pi_1$-высказывания неопровержимо установлена, в то время как в его рассуждениях имеется ошибка, которую он обнаружит лишь значительно позднее. Когда ошибка наконец обнаруживается, математик ясно видит, что его ранние рассуждения неверны, причем в соответствии с теми же самыми критериями, какими он руководствовался и ранее; разница лишь в том, что ранее ошибка замечена не была, — и вот $\Pi_1$-высказывание, полагаемое неопровержимо истинным тогда, воспринимается сейчас как абсолютно ложное (и наоборот).

Мы вполне можем ожидать подобного поведения и от робота, т. е. на его ☆-утверждения, вообще говоря, полагаться нельзя, пусть даже он и удостоил их самолично статуса ☆. Впоследствии робот может исправить свою ошибку, однако ошибка-то уже сделана. Каким образом это обстоятельство отразится на нашем выводе относительно обоснованности формальной системы $\mathbb{Q}(\mathbf{M})$? Очевидно, что система $\mathbb{Q}(\mathbf{M})$ *не является* целиком и полностью обоснованной, не «воспринимает» ее как таковую и робот, так что его гёделевскому предположению $G(\mathbb{Q}(\mathbf{M}))$ доверять нельзя. К этому, в сущности, и сводится суть оговорки (b).

Попробуем выяснить, может ли наш робот, приходя к тому или иному «неопровержимому» заключению, что-либо иметь в виду, и если да, то что именно. Уместно сопоставить эту ситуацию с той, что мы рассматривали в случае математика-человека. Тогда нас не занимало, что конкретно случилось обнаружить какому-либо *реальному* математику, нас занимало лишь то, что может быть принято за неопровержимую истину *в принципе*. Вспомним также знаменитую фразу Фейнмана: «Не слушайте, что я говорю; слушайте, что я имею в виду!». Похоже, нам нет необходимости исследовать то, что робот говорит, исследовать нужно то, что он имеет в виду. Не совсем, впрочем, ясно (особенно если исследователь имеет несчастье являться приверженцем скорее точки зрения $\mathscr{B}$, нежели $\mathscr{A}$), как следует интерпретировать саму идею того, что робот способен что бы то ни было *иметь в виду*. Если бы было возможно опираться не на то, что робот ☆-утверждает, а на то, что он в действительности «имеет в виду», либо на то, что он в принципе «должен иметь в виду», то тогда проблему возможной неточности его ☆-утверждений можно было бы обойти. Беда, однако, в том, что в нашем распоряжении, по всей видимости, нет никаких средств, позволяющих снаружи получить доступ к информации о том, что робот «имеет в виду» или о том, что, «как ему кажется, он имеет в виду». До тех пор, пока речь идет о формальной системе $\mathbb{Q}(\mathbf{M})$, нам, судя по всему, придется полагаться лишь на доступные ☆-утверждения, в достоверности которых мы не можем быть полностью уверены.

Не здесь ли проходит возможная операционная граница между точками зрения $\mathscr{A}$ и $\mathscr{B}$? Не исключено, что так оно и есть; хотя позиции $\mathscr{A}$ и $\mathscr{B}$ эквивалентны в отношении принципиальной возможности внешних проявлений сознательной деятельности в поведении физической системы, люди, этих позиций придерживающиеся, могут разойтись в своих *ожиданиях* как раз в вопросе о том, какую именно вычислительную систему можно рассматривать как способную осуществить эффективное моделирование мозговой активности человека, находящегося в процессе осознания справедливости того или иного математического положения (см. конец §3.12). Как бы то ни было, возможные расхождения в такого рода ожиданиях не имеют к нашему исследованию сколько-нибудь существенного отношения.



## 3.18. Введение случайности: ансамбли всех возможных роботов

В отсутствие прямого операционного метода разрешения этих семантических проблем нам придется полагаться на конкретные ☆-утверждения, которые наш робот будет делать, побуждаемый механизмами, управляющими его поведением. Нам придется смириться с тем, что некоторые из этих утверждений могут оказаться ошибочными, однако такие ошибки исправимы и, в общем случае, чрезвычайно редки. Разумно будет предположить, что всякий раз, когда робот допускает ошибку в одном из своих ☆-утверждений, ошибку эту можно приписать (по меньшей мере частично) каким-то случайным факторам, присутствующим в окружении или во внутренних процедурах робота. Если вообразить себе второго робота, функционирующего в соответствии с механизмами того же типа, что управляют поведением первого робота, однако при участии иных случайных факторов, то этот второй робот вряд ли совершит те же ошибки, что и первый, — но вполне может совершить другие. Упомянутые факторы могут привноситься теми самыми подлинно случайными элементами, которые определяются либо как часть информации, поступающей на вход робота из внешнего окружения, либо как компоненты внутренних процедур робота. Как вариант, они могут представлять собой псевдослучайные результаты неких детерминистских, но хаотических вычислений, как внешних, так и внутренних.

В рамках настоящего рассуждения я буду полагать, что ни один из подобных псевдослучайных элементов не играет в происходящем иной роли, чем та, которую могут выполнить (по меньшей мере с тем же успехом) элементы подлинно случайные. Вполне естественная, на мой взгляд, позиция. Впрочем, не исключается и возможность обнаружения в поведении хаотических систем (отнюдь не сводящемся только лишь к моделированию случайности) чего-то такого, что может послужить приближением какой-либо интересующей нас разновидности невычислительного поведения. Я не припомню, чтобы такая возможность где-либо всерьез обсуждалась, хотя есть люди, которые твердо убеждены в том, что хаотическое поведение представляет собой фундаментальный аспект деятельности мозга. Лично для меня подобные аргументы останутся неубедительными до тех пор, пока мне не продемонстрируют какое-нибудь существенно

*не*случайное (т. е. непсевдослучайное) поведение такой хаотической системы — поведение, которое может в сколько-нибудь сильном смысле являться приближением поведения подлинно невычислительного. Ни один намек на подобного рода демонстрацию моих ушей пока не достиг. Более того, как мы подчеркнем несколько позднее (§ 3.22), в любом случае маловероятно, что хаотическое поведение сможет проигнорировать те сложности, которые представляет для вычислительной модели разума гёделевское доказательство.

Допустим пока, что любые псевдослучайные (или иным образом хаотические) элементы в поведении нашего робота или в его окружении можно заменить элементами подлинно случайными, причем без какой бы то ни было потери эффективности. Для выяснения роли подлинной случайности нам необходимо составить *ансамбль* из всех возможных альтернативных вариантов. Поскольку мы предполагаем, что наш робот имеет цифровое управление, и, соответственно, его окружение также можно реализовать в каком-либо цифровом виде (вспомним о «внутренних» и «внешних» участках ленты нашей описанной выше машины Тьюринга; см. также § 1.8), то количество подобных возможных альтернатив непременно будет *конечным*. Это число может быть *очень* большим, и все же полное описание всех упомянутых альтернатив представляет собой задачу чисто вычислительного характера. Таким образом, и сам полный ансамбль всех возможных роботов, каждый из которых действует в соответствии с заложенными нами механизмами, составляет всего-навсего вычислительную систему — пусть даже такую, какую нам вряд ли удастся реализовать на практике, используя те компьютеры, которыми мы располагаем в настоящее время или можем вообразить в обозримом будущем. Тем не менее, несмотря на малую вероятность практического осуществления совокупного моделирования всех возможных роботов, функционирующих в соответствии с набором механизмов **M**, само вычисление «непознаваемым» считаться не может; иначе говоря, мы способны понять (теоретически), как построить такой компьютер — или машину Тьюринга, — который с подобным моделированием справится, пусть даже оно пока и не осуществимо *практически*. В этом состоит ключевой момент нашего рассуждения. Познаваемым механизмом или познаваемым вычислением является тот механизм или то вычисление, которое человек способен *описать*; совсем



не обязательно действительно выполнять это вычисление ни самому человеку, ни даже компьютеру, который человек в состоянии в данных обстоятельствах построить. Ранее (в комментарии к **Q8**) мы уже высказывали весьма похожее соображение; и то, и другое вполне согласуются с терминологией, введенной в начале § 3.5.

## 3.19. Исключение ошибочных ☆-утверждений

Вернемся к вопросу об ошибочных (но допускающих исправление) ☆-утверждениях, которые может время от времени выдавать наш робот. Предположим, что робот такую ошибку все-таки совершил. Если мы можем допустить, что какой-либо другой робот, или тот же робот несколько позднее, или другой *экземпляр* того же робота такую же ошибку вряд ли совершит, то мы *в принципе* сможем установить факт ошибочности данного ☆-утверждения, проанализировав действия ансамбля из всех возможных роботов. Представим себе, что моделирование поведения всей совокупности возможных роботов осуществляется в нашем случае таким образом, что различные этапы развития различных экземпляров нашего робота мы рассматриваем как одновременные. (Это делается лишь для удобства рассмотрения и никоим образом не подразумевает, что для такого моделирования непременно требуется параллельное выполнение действий. Как мы уже видели, принципиальных различий, помимо эффективности, между параллельным и последовательным выполнением вычислений нет; см. § 1.5). Такой подход должен, в принципе, дать нам возможность уже на стадии рассмотрения результата моделирования выделить из общей массы корректных ☆-утверждений редкие (относительно) ошибочные ☆-утверждения, воспользовавшись тем обстоятельством, что ошибочные утверждения «исправимы» и будут посему однозначно идентифицироваться как ошибочные подавляющим большинством участвующих в модели экземпляров нашего робота, — по крайней мере, с накоплением с течением времени (модельного) различными экземплярами робота достаточного параллельного «опыта». Я вовсе не требую, чтобы подобная процедура была осуществима на практике; достаточно, чтобы она была вычислительной, а лежащие в основе всего этого вычисления *правила* **M** — в принципе «познаваемыми».

Для того чтобы приблизить нашу модель к виду, приличествующему человеческому математическому сообществу, а также лишний раз удостовериться в отсутствии ошибок в ☆-утверждениях, рассмотрим ситуацию, в которой все окружение нашего робота разделяется на две части: *сообщество* других роботов и остальное, лишенное роботов (а также и людей), окружение; в дополнение к остальному окружению, в модель следует ввести некоторое количество учителей, по крайней мере, на ранних этапах развития роботов, и хотя бы для того, чтобы все роботы одинаково понимали строгий смысл присвоения тому или иному утверждению статуса ☆. В моделируемый нами ансамбль войдут на правах различных экземпляров все возможные различные варианты поведения *всех* роботов, а также все возможные (релевантные) варианты остального окружения и предоставляемых человеком сведений, варьирующиеся в зависимости от конкретного выбора задействованных в модели случайных параметров. Как и ранее, правила, по которым будет функционировать наша модель (и которые я опять обозначу буквой **M**), можно полагать в полной мере познаваемыми, невзирая на необычайную сложность всех сопутствующих расчетов, необходимых для ее практической реализации.

Предположим, что мы берем на заметку все (в принципе) $\Pi_1$-высказывания, ☆-утверждаемые (а также все высказывания с ☆-утвержденными отрицаниями) любым из всевозможных экземпляров наших (вычислительно моделируемых) роботов. Объединим все подобные ☆-утверждения в отдельную группу и назовем их *безошибочными*. Далее, мы можем потребовать, чтобы любое ☆-утверждение относительно того или иного $\Pi_1$-высказывания *игнорировалось*, если в течение некоторого промежутка времени $T$ (в прошлом или в будущем) количество $r$ различных экземпляров этого ☆-утверждения в ансамбле из всех одновременно действующих роботов не удовлетворит неравенству $r > L + Ns$, где $L$ и $N$ суть некоторые достаточно большие числа, а $s$ — количество ☆-утверждений, производимых в течение того же промежутка времени и занимающих относительно рассматриваемого $\Pi_1$-высказывания противоположную позицию либо просто утверждающих, что рассуждения, на которые опирается исходное ☆-утверждение, ошибочны. При желании мы можем настаивать на том, чтобы промежуток времени $T$ (это время не обязательно должно совпадать с «реальным» моделируемым временем и может измеряться в некоторых единицах вычислительной активности), равно как и числа $L$ и $N$, увеличивался по мере увеличения «сложности» ☆-утверждаемого $\Pi_1$-высказывания.



Понятию «сложности» применительно к $\Pi_1$-высказываниям можно придать точный характер на основании спецификаций машины Тьюринга, как мы это уже делали в § 2.6 (в конце комментария к возражению **Q8**). Для большей конкретности мы можем воспользоваться явными формулировками, представленными в НРК (глава 2), как вкратце показано в приложении А (а это уже здесь, с. 193). Итак, *степенью сложности* $\Pi_1$-высказывания, утверждающего незавершаемость вычисления $T_m(n)$ машины Тьюринга, мы будем полагать число $\rho$ знаков в двоичном представлении *большего* из пары чисел $m$ и $n$.

Причина введения в данное рассуждение числа $L$ — вместо того чтобы удовлетвориться какой-нибудь огромной величиной в лице одного лишь коэффициента $N$, — заключается в необходимости учета следующей возможности. Предположим, что внутри нашего ансамбля, благодаря редчайшей случайности, появляется «безумный» робот, который формулирует какое-нибудь абсолютно нелепое ☆-утверждение, ничего не сообщая о нем остальным роботам, причем нелепость этого утверждения настолько велика, что ни одному из роботов никогда не придет в «голову» — хотя бы просто на всякий случай — сформулировать его опровержение. В отсутствие числа $L$ такое ☆-утверждение автоматически попадет, в соответствии с нашими критериями, в группу «безошибочных». Введение же достаточно большого $L$ такую ситуацию предотвратит — при условии, разумеется, что подобное «безумие» возникает среди роботов не часто. (Вполне возможно, что я упустил из виду еще что-нибудь, и необходимо будет позаботиться о каких-то дополнительных мерах предосторожности. Представляется разумным, однако, по крайней мере на данный момент, ограничиться критериями, предложенными выше.)

Учитывая, что все ☆-утверждения, согласно исходному допущению, следует полагать «неопровержимыми» заявлениями нашего робота (основанными на, по всей видимости, присущих роботу четких логических принципах и посему не содержащими ничего такого, в чем робот испытывает хотя бы малейшее сомнение), то вполне разумным представляется предположение, что

вышеописанным образом действительно можно устранить редкие промахи в рассуждениях робота, причем функции $T(\rho)$, $L(\rho)$ и $N(\rho)$ вряд ли окажутся чем-то из ряда вон выходящим. Предположив, что все так и есть, мы опять получаем не что иное, как *вычислительную* систему — систему *познаваемую* (в том смысле, что познаваемыми являются лежащие в основе системы *правила*) при условии познаваемости исходного набора механизмов **M**, определяющего поведение нашего робота. Эта вычислительная система дает нам новую формальную систему $\mathbb{Q}'(\mathbf{M})$ (также познаваемую), теоремами которой являются те самые *безошибочные* ☆-утверждения (либо утверждения, выводимые из них посредством простых логических операций исчисления предикатов).

Вообще говоря, для нас с вами важно не столько то, что эти утверждения *действительно* безошибочны, сколько то, что в их безошибочности *убеждены* сами роботы (для приверженцев точки зрения $\mathscr{B}$ особо оговоримся, что концепцию роботовой «убежденности» следует понимать в чисто операционном смысле *моделирования* роботом этой самой убежденности, см. §§ 3.12, 3.17).

Если точнее, то нам требуется, чтобы робот был готов поверить в то, что упомянутые ☆-утверждения действительно безошибочны, *исходя из допущения*, что именно набором механизмов **M** и определяется его поведение (гипотеза $\mathscr{M}$ из § 3.16). До сих пор, в данном разделе, мы занимались исключительно устранением ошибок в ☆-утверждениях робота. Однако, *на самом деле*, ввиду представленного в § 3.16 фундаментального противоречия, нас интересует устранение ошибок в его $☆_{\mathscr{M}}$-утверждениях, т. е. в тех $\Pi_1$-высказываниях, что по неопровержимой убежденности робота следуют из гипотезы $\mathscr{M}$. Поскольку принятие роботами формальной системы $\mathbb{Q}'(\mathbf{M})$ в любом случае обусловлено гипотезой $\mathscr{M}$, мы вполне можем предложить им для обдумывания и более обширную формальную систему $\mathbb{Q}'_{\mathscr{M}}(\mathbf{M})$, определяемую аналогично формальной системе $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ из § 3.16. Под $\mathbb{Q}'_{\mathscr{M}}(\mathbf{M})$ в данном случае понимается формальная система, построенная из $☆_{\mathscr{M}}$-утверждений, «безошибочность» которых установлена в соответствии с вышеописанными критериями $T$, $L$ и $N$. В частности, утверждение «утверждение $G(\mathbb{Q}'_{\mathscr{M}}(\mathbf{M}))$ истинно» считается здесь безошибочным $☆_{\mathscr{M}}$-утверждением. Те же рассуждения, что и в § 3.16, приводят нас к выводу, что роботы не смогут принять допущение, что они построены в соответствии с набором механизмов **M** (вкупе с проверочными критериями $T$, $L$ и $N$), независимо от того, какие именно вычислительные правила **M** мы им предложим.



Достаточно ли этих соображений для того, чтобы окончательно удостовериться в наличии противоречия? У читателя, возможно, осталось некое тревожное ощущение — кто знает, вдруг сквозь тщательно расставленные сети, невзирая на все наши старания, проскользнули какие-нибудь ошибочные $☆_{\mathscr{M}}$- или ☆-утверждения? В конце концов, приведенные выше рассуждения будут иметь смысл лишь в том случае, если нам удастся исключить абсолютно *все* ошибочные $☆_{\mathscr{M}}$-утверждения (или ☆-утверждения) в отношении $\Pi_1$-высказываний. Окончательно и бесповоротно *удостовериться* в истинности утверждения $G(\mathbb{Q}'_{\mathscr{M}}(\mathbf{M}))$ нам (и роботам) поможет *обоснованность* формальной системы $\mathbb{Q}'_{\mathscr{M}}(\mathbf{M})$ (обусловленная гипотезой $\mathscr{M}$). Эта самая обоснованность подразумевает, что система $\mathbb{Q}'_{\mathscr{M}}(\mathbf{M})$ *ни в коем случае* не может содержать таких $☆_{\mathscr{M}}$-утверждений, которые являются — или всего лишь предполагаются — ошибочными. Невзирая на все предпринятые меры предосторожности, полной уверенности у нас (да и у роботов, полагаю) все-таки нет — хотя бы по той простой причине, что количество возможных утверждений подобного рода бесконечно.

## 3.20. Возможность ограничиться конечным числом $☆_{\mathscr{M}}$-утверждений

Есть, впрочем, возможность именно эту конкретную проблему разрешить и сузить область рассмотрения до *конечного* множества различных $☆_{\mathscr{M}}$-утверждений. Само доказательство несколько громоздко, однако основная идея заключается в том, что следует рассматривать только те $\Pi_1$-высказывания, спецификации которых являются «краткими» в некотором вполне определенном смысле. Конкретная степень необходимой «краткости» зависит от того, насколько сложное описание системы механизмов **M** нам необходимо. Чем сложнее описание **M**, тем «длиннее» допускаемые к рассмотрению $\Pi_1$-высказывания. «Максимальная длина» задается неким числом $c$, которое можно

определить из степени сложности правил, определяющих формальную систему $\mathbb{Q}'_{\mathcal{M}}(\mathbf{M})$. Смысл в том, что при переходе к гёделевскому предположению для этой формальной системы — которую нам, вообще говоря, придется слегка модифицировать — мы получим утверждение, сложность которого будет лишь немногим выше, нежели сложность такой модифицированной системы. Таким образом, проявив должную осторожность при выборе числа $c$, мы можем добиться того, что и гёделевское предположение будет также «кратким». Это позволит нам получить требуемое противоречие, не выходя за пределы конечного множества «кратких» $\Pi_1$-высказываний.

Подробнее о том, как это осуществить на практике, мы поговорим в оставшейся части настоящего раздела. Тем из читателей, кого такие подробности не занимают (уверен, таких наберется немало), я рекомендую просто-напросто пропустить весь этот материал.

Нам понадобится несколько модифицировать формальную систему $\mathbb{Q}'_{\mathcal{M}}(\mathbf{M})$, приведя ее к виду $\mathbb{Q}'_{\mathcal{M}}(\mathbf{M}, c)$ — для краткости я буду обозначать ее просто как $\mathbb{Q}(c)$ (отброшенные обозначения в данной ситуации несущественны и лишь добавляют путаницы и громоздкости). Формальная система $\mathbb{Q}(c)$ определяется следующим образом: при построении этой системы допускается принимать в качестве «безошибочных» только те $\bigstar_{\mathcal{M}}$-утверждения, степень сложности которых (задаваемая описанным выше числом $\rho$) меньше $c$, где $c$ есть некоторое должным образом выбранное число, подробнее о котором я расскажу чуть ниже. Для «безошибочных» $\bigstar_{\mathcal{M}}$-утверждений, удовлетворяющих неравенству $\rho < c$, я буду использовать обозначение «$\surd$краткие $\bigstar_{\mathcal{M}}$-утверждения». Как и прежде, множество действительных *теорем* формальной системы $\mathbb{Q}(c)$ будет включать в себя не только $\surd$краткие $\bigstar_{\mathcal{M}}$-утверждения, но также и утверждения, получаемые из $\surd$кратких $\bigstar_{\mathcal{M}}$-утверждений посредством стандартных логических операций (позаимствованных, скажем, из исчисления предикатов). Хотя количество теорем системы $\mathbb{Q}(c)$ бесконечно, все они выводятся с помощью обыкновенных логических операций из *конечного* множества $\surd$кратких $\bigstar_{\mathcal{M}}$-утверждений. Далее, поскольку мы ограничиваем рассмотрение конечным множеством, мы вполне можем допустить, что функции $T$, $L$ и $N$ *постоянны* (и принимают, скажем, наибольшие значения на конечном интервале $\rho$). Таким образом, формальная система $\mathbb{Q}(c)$ задается



лишь четырьмя постоянными $c$, $T$, $L$, $N$ и общей системой механизмов $\mathbf{M}$, определяющих поведение робота.

Отметим существенный для наших рассуждений момент: гёделевская процедура строго *фиксирована* и не нуждается в увеличении сложности выше некоторого определенного предела. Гёделевским предположением $G(\mathbb{H})$ для формальной системы $\mathbb{H}$ является $\Pi_1$-высказывание, степень сложности которого должна лишь на сравнительно малую величину превышать степень сложности самой системы $\mathbb{H}$, причем эту величину можно определить точно.

Конкретности ради я позволю себе некоторое нарушение системы обозначений и буду вкладывать в запись «$G(\mathbb{H})$» некий особый смысл, который может и не совпасть в точности с определением, данным в §2.8. В формальной системе $\mathbb{H}$ нас интересует лишь ее способность доказывать $\Pi_1$-высказывания. В силу этой своей способности система $\mathbb{H}$ дает нам алгебраическую процедуру $A$, с помощью которой мы можем в точности установить (на основании завершения выполнения $A$) справедливость тех $\Pi_1$-высказываний, формулировка которых допускается правилами системы $\mathbb{H}$. А под $\Pi_1$-высказыванием понимается утверждение вида «действие машины Тьюринга $T_p(q)$ не завершается» — здесь и далее мы будем пользоваться специальным способом маркировки машин Тьюринга, описанным в Приложении А (или в НРК, глава 2). Мы полагаем, что процедура $A$ выполняется над парой чисел $(p, q)$, как в §2.5. Таким образом, собственно вычисление $A(p, q)$ завершается в том и *только* в том случае, если в рамках формальной системы $\mathbb{H}$ возможно установить справедливость того самого $\Pi_1$-высказывания, которое утверждает, что «действие $T_p(q)$ *не* завершается». С помощью описанной в §2.5 процедуры мы получили некое конкретное вычисление (обозначенное там как «$C_k(k)$»), а вместе с ним, при условии обоснованности системы $\mathbb{H}$, и истинное $\Pi_1$-высказывание, которое системе $\mathbb{H}$ оказывается «не по зубам». *Именно* это $\Pi_1$-высказывание я буду теперь обозначать через $G(\mathbb{H})$. Оно существенно эквивалентно (при условии достаточной обширности $\mathbb{H}$) действительному утверждению «система $\mathbb{H}$ непротиворечива», хотя в некоторых деталях эти два утверждения могут и не совпадать (см. §2.8).

Пусть $\alpha$ есть *степень сложности* процедуры $A$ (по определению, данному в §2.6, в конце комментария к возражению **Q8**) — иными словами, количество знаков в двоичном представлении

числа $a$, где $A = T_a$. Тогда, согласно построению, представленному в явном виде в Приложении А, находим, что степень сложности $\eta$ утверждения $G(\mathbb{H})$ удовлетворяет неравенству $\eta < \alpha + 210\log_2(\alpha + 336)$. Для нужд настоящего рассуждения мы можем определить степень сложности формальной системы $\mathbb{H}$ как равную степени сложности процедуры $A$, т. е. числу $\alpha$. Приняв такое определение, мы видим, что «излишек» сложности, связанный с переходом от $\mathbb{H}$ к $G(\mathbb{H})$, оказывается еще меньше, чем и без того относительно крохотная величина $210\log_2(\alpha + 336)$.

Далее нам предстоит показать, что если $\mathbb{H} = \mathbb{Q}(c)$ при достаточно большом $c$, то $\eta < c$. Отсюда, соответственно, последует, что и $\Pi_1$-высказывание $G(\mathbb{Q}(c))$ должно оказаться в пределах досягаемости системы $\mathbb{Q}(c)$ при условии, что роботы принимают $G(\mathbb{Q}(c))$ с $\bigstar$-убежденностью. Доказав, что $c > \gamma + 210\log_2(\gamma + 336)$, мы докажем и то, что $\gamma < c$; буквой $\gamma$ мы обозначили значение $\alpha$ при $\mathbb{H} = \mathbb{Q}(c)$. Единственная возможная сложность здесь обусловлена тем обстоятельством, что сама величина $\gamma$ зависит от $c$, хотя и не обязательно очень сильно. Эта зависимость $\gamma$ от $c$ имеет две различных причины. Во-первых, число $c$ являет собой явный предел степени сложности тех $\Pi_1$-высказываний, которые в определении формальной системы $\mathbb{Q}(c)$ называются «безошибочными $\bigstar_{\mathscr{M}}$-утверждениями»; вторая же причина происходит из того факта, что система $\mathbb{Q}(c)$ явным образом обусловлена выбором чисел $T$, $L$ и $N$, и можно предположить, что для принятия в качестве «безошибочного» $\bigstar_{\mathscr{M}}$-утверждения большей сложности необходимы какие-то более жесткие критерии.

Относительно первой причины зависимости $\gamma$ от $c$ отметим, что описание действительной величины числа $c$ необходимо задавать в явном виде только однажды (после чего внутри системы достаточно обозначения $c$). Если при задании величины $c$ используется чисто двоичное представление, то (при больших $c$) такое описание дает всего-навсего логарифмическую зависимость $\gamma$ от $c$ (поскольку количество знаков в двоичном представлении натурального $n$ равно приблизительно $\log_2 n$). Вообще говоря, учитывая, что число $c$ интересует нас лишь в качестве возможного предела, точное значение которого находить вовсе не обязательно, мы можем поступить гораздо более остроумным образом. Например, число $2^{2^{\cdot^{\cdot^{2}}}}$ с $s$ показателями можно задать с помощью $s$



символов или около того, и вовсе нетрудно подыскать примеры, в которых величина задаваемого числа возрастает с ростом $s$ еще быстрее. Сгодится любая вычислимая функция от $s$. Иными словами, для того чтобы задать предел $c$ (при достаточно большом значении $c$), необходимо всего лишь несколько символов.

Что касается второй причины, т. е. зависимости от $c$ чисел $T$, $L$ и $N$, то, в силу вышеизложенных соображений, представляется очевидным, что для задания величин этих чисел (в особенности, их возможных предельных значений) совершенно не требуется, чтобы количество знаков в их двоичном представлении возрастало так же быстро, как $c$; более чем достаточно будет и, скажем, обыкновенной логарифмической зависимости от $c$. Следовательно, мы с легкостью можем допустить, что зависимость величины $\gamma + 210\log_2(\gamma + 336)$ от $c$ является не более чем грубо логарифмической, а также устроить так, чтобы само число $c$ всегда было больше этой величины.

Согласимся с таким выбором $c$ и будем в дальнейшем вместо $\mathbb{Q}(c)$ записывать $\mathbb{Q}^*$. Итак, $\mathbb{Q}^*$ есть формальная система, теоремами которой являются все математические высказывания, какие можно вывести из конечного количества $\sqrt{}$кратких $\bigstar_{\mathscr{M}}$-утверждений, используя стандартные логические правила (исчисление предикатов). Количество этих $\bigstar_{\mathscr{M}}$-утверждений конечно, поэтому разумным будет предположить, что для гарантии их действительной безошибочности вполне достаточно некоторого набора постоянных $T$, $L$ и $N$. Если роботы верят в это с $\bigstar_{\mathscr{M}}$-убежденностью, то они, несомненно, $\bigstar_{\mathscr{M}}$-заключат, что гёделевское предположение $G(\mathbb{Q}^*)$ также истинно на основании гипотезы $\mathscr{M}$, поскольку является $\Pi_1$-высказыванием меньшей, нежели $c$, сложности. Рассуждение для получения утверждения $G(\mathbb{Q}^*)$ из $\bigstar_{\mathscr{M}}$-убежденности в обоснованности формальной системы $\mathbb{Q}^*$ достаточно просто (в сущности, я его уже привел), так что с присвоением этому утверждению статуса $\bigstar_{\mathscr{M}}$ проблем возникнуть не должно. То есть само $G(\mathbb{Q}^*)$ также должно быть теоремой системы $\mathbb{Q}^*$. Это, однако, противоречит убежденности роботов в обоснованности $\mathbb{Q}^*$. Таким образом, упомянутая убежденность (при условии справедливости гипотезы $\mathscr{M}$ и достаточно больших числах $T$, $L$ и $N$) оказывается несовместимой с убежденностью в том, что поведением роботов действительно управляют механизмы **M**, — а значит, механизмы **M** поведением роботов управлять *не могут*.

Как же роботы могут удостовериться в том, что были выбраны достаточно большие числа $T$, $L$ и $N$? Никак. Вместо этого они могут выбрать *некоторый* набор таких чисел и попробовать допустить, что те достаточно велики, — и прийти в результате к противоречию с исходным предположением, согласно которому их поведение обусловлено набором механизмов **M**. Далее они вольны предположить, что достаточным окажется набор из несколько бóльших чисел, — снова прийти к противоречию и т. д. Вскоре они сообразят, что к противоречию они приходят при *любом* выборе значений (вообще говоря, здесь нужно учесть, помимо прочего, небольшой технический момент, суть которого состоит в том, что при совершенно уже запредельных значениях $T$, $L$ и $N$ значение $c$ также должно будет несколько подрасти — однако это неважно). Таким образом, получая один и тот же результат вне зависимости от значений $T$, $L$ и $N$, роботы — равно как, по всей видимости, и мы — приходят к заключению, что в основе их математических мыслительных процессов не может лежать познаваемая вычислительная процедура **M**, *какой бы она ни была*.

## 3.21. Окончателен ли приговор?

Отметим, что к такому же выводу мы придем и в случае принятия нами самых разных возможных мер предосторожности, причем вовсе необязательно подобных тем, что я предлагал выше. Наверняка в предложенную модель можно еще внести множество усовершенствований. Можно, например, предположить, что роботы в результате длительной работы впадают в «старческое слабоумие», их сообщества вырождаются, а стандарты падают, т. е. увеличение числа $T$ выше определенного значения на деле *увеличивает* и вероятность ошибки в $\text{☆}_{\mathcal{M}}$-утверждениях. С другой стороны, если слишком большим сделать $N$ (или $L$), то возникает риск исключить вообще все $\text{☆}_{\mathcal{M}}$-утверждения из-за существующего в сообществе меньшинства «глупых» роботов, разражающихся время от времени произвольными $\text{☆}_{\mathcal{M}}$-утверждениями, которые в данном случае не перекроются необходимым количеством ☆-утверждений, формулируемых роботами здравомыслящими. Несомненно, не составит большого труда такой риск полностью исключить, введя еще несколько ограничивающих параметров или, скажем, сформировав группу элитных роботов, силами которых рядовые члены сообщества будут непрерывно тестироваться на предмет адекватности своих интеллектуальных способностей, и потребовав к тому же, чтобы статус ☆ присваивался утверждениям только с одобрения всего сообщества роботов в целом.



Существует и много других возможностей улучшения качества $\text{☆}_{\mathcal{M}}$-утверждений или исключения ошибочных утверждений из общего (конечного) их числа. Кого-то, возможно, обеспокоит тот факт, что, несмотря на установление предела $c$ сложности $\Pi_1$-высказываний, ограничивающего общее количество кандидатов на ☆- или $\text{☆}_{\mathcal{M}}$-статус до некоторой конечной величины, эта величина окажется все же чрезвычайно огромной (будучи экспоненциально зависимой от $c$), вследствие чего становится весьма сложно *однозначно* удостовериться, что исключены *все* возможные ошибочные $\text{☆}_{\mathcal{M}}$-утверждения. В самом деле, никакого ограничения не задается в рамках нашей модели на количество «робото-вычислений», необходимых для получения удовлетворительного $\text{☆}_{\mathcal{M}}$-доказательства какого-либо из $\Pi_1$-высказываний. Следует ввести четкое правило: чем длиннее в таком доказательстве цепь рассуждений, тем более жесткие критерии применяются при решении вопроса о присвоении ему $\text{☆}_{\mathcal{M}}$-статуса. В конце концов, математики-люди реагировали бы именно так. Прежде чем принять в качестве неопровержимого доказательства собрание многочисленных путаных аргументов, мы, естественно, чрезвычайно долго и придирчиво его изучаем. Аналогичные соображения, разумеется, применимы и к тому случаю, когда предложенное доказательство на предмет его соответствия $\text{☆}_{\mathcal{M}}$-статусу исследуют роботы.

Вышеприведенные рассуждения в равной степени справедливы и в случае любой дальнейшей модификации условий, имеющих целью устранение ошибок, при условии, что характер такой модификации в некоем широком смысле аналогичен характеру уже предложенных. Для того чтобы эти рассуждения работали, необходимо лишь наличие *какого угодно* четко сформулированного и вычислимого условия, достаточного для устранения всех ошибочных $\text{☆}_{\mathcal{M}}$-утверждений. В результате мы приходим к строгому выводу: *никакие познаваемые механизмы, пусть и снабженные какими угодно вычислительными «подпорка-*

*ми», не способны воспроизвести корректное математическое умозаключение человека.*

Мы рассматривали $\bigstar_{\mathcal{M}}$-утверждения, которые, оказавшись по той или иной причине ошибочными, в принципе *исправимы* самими роботами, — пусть даже в каком-то конкретном экземпляре модели роботова сообщества эти утверждения так и остаются неисправленными. Что же еще может означать (в операционном смысле) фраза «в принципе исправимы», как не «исправимы средствами некоторой общей процедуры, подобной тем, что предложены выше»? Ошибка, которую не исправил позднее тот робот, что ее допустил, может быть исправлена каким-либо другим роботом — более того, большинство потенциально существующих экземпляров первого робота эту конкретную ошибку вообще не допустят. Делаем вывод (с одной, по-видимому, незначительной оговоркой, суть которой в том, что хаотические компоненты нашей модели можно еще заменить на подлинно случайные; см. ниже, § 3.22): никакой набор познаваемых вычислительных правил **M** (неизменных нисходящих, «самосовершенствующихся» восходящих либо и тех, и других в какой угодно пропорции) не может обусловливать поведение нашего сообщества роботов, равно как и отдельных его членов, — *если* исходить из допущения, что роботы способны достичь человеческого уровня математического понимания. Вообразив, что мы сами функционируем как управляемые вычислительными правилами роботы, мы оказываемся перед непреодолимым противоречием.

## 3.22. Спасет ли вычислительную модель разума хаос?

Вернемся ненадолго к вопросу о хаосе. Хотя, как неоднократно подчеркивается в этой книге (в частности, в § 1.7), хаотические системы в том виде, в каком они обычно рассматриваются, представляют собой всего-навсего особого рода вычислительные системы, довольно широко распространено мнение о том, что феномен хаоса может иметь весьма значительное отношение к деятельности мозга. В представленных выше рассуждениях я опирался, с одной стороны, на обоснованное, как мне кажется, предположение, согласно которому любое хаотическое вычислительное поведение можно без существенной потери функциональности заменить поведением подлинно случайным. Против такого



допущения можно привести, по крайней мере, одно вполне оправданное возражение. Поведение хаотической системы — пусть мы и ожидаем от него огромной сложности в мельчайших деталях и *видимой* случайности — *в действительности* случайным не является. В самом деле, многие хаотические системы демонстрируют весьма интересное сложное поведение, явно отклоняющееся от чистой случайности. (Иногда для описания сложного неслучайного поведения[10], демонстрируемого хаотическими системами, используется термин «край хаоса».) Возможно ли, чтобы именно в *хаосе* крылась разгадка тайны человеческого интеллекта? Если это так, то нам предстоит понять нечто доселе абсолютно неведомое относительно того, как ведут себя в соответствующих ситуациях хаотические системы. Хаотической системе в такой ситуации придется очень близко аппроксимировать *невычислительное поведение* в асимптотическом пределе — или нечто подобное. Демонстрации такого поведения, насколько мне известно, еще никто не представлял. Возможность, тем не менее, интересная, и я надеюсь, что в последующие годы ею кто-нибудь всерьез займется.

И все же, безотносительно к упомянутой возможности, хаос может предоставить нам лишь очень сомнительный способ обойти неутешительное заключение, к которому мы пришли в предыдущем параграфе. В представленных выше рассуждениях эффективная хаотическая неслучайность (т. е. непсевдослучайность) играла хоть какую-то роль один-единственный раз — когда мы рассматривали моделирование не просто «действительного» поведения нашего робота (или сообщества роботов), но полный ансамбль всех *возможных* действий роботов, согласующихся с заданным набором механизмов **M**. Та же аргументация применима и здесь, только на сей раз мы не станем включать в эту случайность хаотические результаты функционирования упомянутых механизмов. Впрочем, некоторые случайные элементы (например, в составе исходных данных, определяющих начальное состояние модели) присутствовать все же могут, а чтобы оперировать *этой* случайностью, мы можем вновь воспользоваться идеей ансамбля и тем самым получить возможность рассмотреть в процессе синхронного моделирования большое количество возможных альтернативных робото-историй. Однако *само* хаотическое поведение нам просто-напросто придется *вычислять* — в чем нет ничего странного: на практике, в математических при-

мерах, хаотическое поведение обыкновенно и вычисляется на компьютере. Ансамбль возможных альтернатив окажется в данном случае не таким большим, каким он мог бы быть, допусти мы аппроксимацию хаоса случайностью. Однако в том случае ансамбль подобного размера был нужен лишь для того, чтобы мы могли лишний раз удостовериться в том, что устранили все возможные ошибки в $\bigstar_{\mathcal{M}}$-утверждениях роботов. Даже если ансамбль включает в себя всего *одну* «историческую линию» сообщества роботов, можно быть совершенно уверенным в том, что при достаточно жестком наборе критериев для присвоения $\bigstar_{\mathcal{M}}$-статуса такие ошибки будут очень быстро устраняться либо самими их виновниками, либо какими-то другими роботами сообщества. В ансамбле умеренного размера, составленном из подлинно случайных элементов, устранение ошибок будет происходить более эффективно, при дальнейшем же расширении ансамбля посредством введения в него случайных аппроксимаций на замену подлинно хаотическому поведению сколько-нибудь существенного роста эффективности не предвидится. Вывод: хаос не избавит нас от проблем, связанных с созданием вычислительной модели разума.

## 3.23. *Reductio ad absurdum* — воображаемый диалог

Многие из представленных в предыдущих разделах рассуждений, мягко говоря, несколько запутаны. Для прояснения ситуации читателю предлагается в качестве этакого резюме воображаемый разговор, состоявшийся в далеком будущем между неким гипотетическим, весьма преуспевающим прикладным специалистом в области ИИ и одним из его наиболее удачных кибернетических созданий. Написан диалог с позиции сильного ИИ. [Примечание: процедура **Q** в повествовании выступает в роли алгоритма $A$ из §2.5, а утверждение $G(\mathbf{Q})$ — в роли незавершающегося вычисления $C_k(k)$. То есть к чтению нижеследующего материала можно переходить сразу после §2.5 без какого бы то ни было ущерба для понимания.]

*Альберт Император имел все основания быть удовлетворенным результатом трудов всей своей жизни. Процедуры, которые он запустил в действие много лет назад, наконец принесли плоды. И вот перед вами точный*



*протокол его беседы с одним из наиболее впечатляющих его творений — роботом выдающихся и потенциально сверхчеловеческих математических способностей по имени Математический Интеллектуальный Киберкомплекс (см. рис. 3.2). Обучение робота почти завершено.*

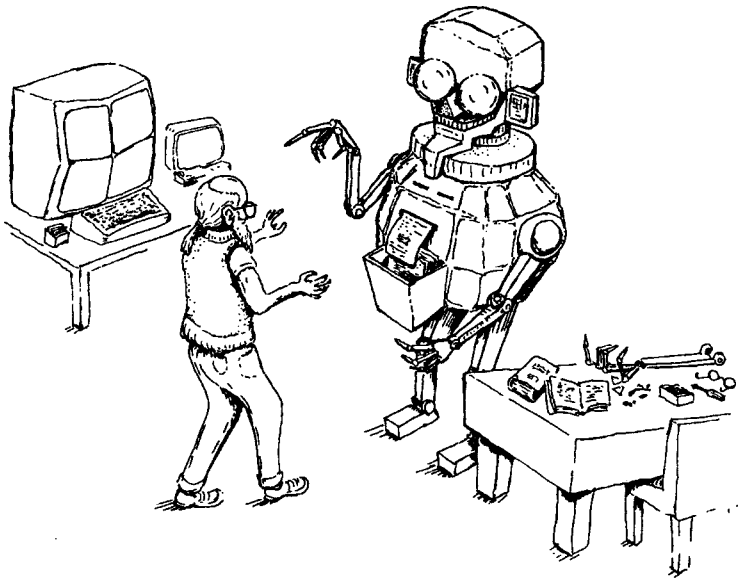


Рис. 3.2. Альберт Император и Математический Интеллектуальный Киберкомплекс.

**Альберт Император:** Просмотрел ли ты статьи, что я давал тебе, — статьи Гёделя, а также и другие, где рассматриваются следствия из его теоремы?

**Математический Интеллектуальный Киберкомплекс:** Разумеется, причем они оказались даже интересными, хотя и довольно элементарными. Этот ваш Гёдель был, по всей видимости, весьма способным логиком... для человека.

**А. И.:** Всего лишь «весьма способным»? Да он был, несомненно, одним из величайших логиков всех времен. Возможно, даже *первым* из величайших!

**М. И. К.:** Приношу извинения, я вовсе не намеревался преуменьшать его заслуги. Вам, разумеется, хорошо известно, что я обучен проявлять общее уважение к достижениям людей (по причине того, что люди очень обидчивы), хотя все эти достижения нам, роботам, обыкновенно представляются весьма тривиальными. Мне просто показалось, что уж с тобой-то я могу, по крайней мере, выражать свои суждения просто и открыто.

**А. И.:** Безусловно, можешь. Прости и ты меня, я был неправ. Так, значит, у тебя не возникло никаких трудностей с пониманием теоремы Гёделя?

**М. И. К.:** Абсолютно никаких. Уверен, я бы и сам додумался до такой теоремы, если бы у меня было хоть немного больше свободного времени. Но мой разум был занят иными, чрезвычайно увлекательными вопросами, связанными с трансфинитной нелинейной когомологией, которая в последнее время интересует меня гораздо больше. Теорема Гёделя показалась мне очень здравой и непосредственной. Повторюсь, совершенно никаких трудностей у меня с ней не возникло.

**А. И.:** А вот получи-ка, Пенроуз!

**М. И. К.:** Пенроуз? Кто такой Пенроуз?

**А. И.:** Да я тут недавно наткнулся на одну старую книжку. Ничего особенного, не стоило и упоминать. Автор, насколько я помню, утверждал, что то, о чем ты мне сейчас рассказал, принципиально невозможно.

**М. И. К.:** Ха-ха-ха! (*Робот поразительно похоже имитирует презрительный смех.*)

**А. И.:** Кстати, эта книжка мне кое о чем напомнила. Показывал ли я тебе когда-нибудь в полном объеме те правила, что мы применили при составлении вычислительных процедур, которые позволили в конечном счете разработать и построить тебя и твоих коллег-роботов?

**М. И. К.:** Нет, пока еще нет. Я надеялся, что когда-нибудь ты все же сделаешь это, и еще я думал, что ты, может быть, полагаешь подробное описание этих процедур чем-то вроде коммерческой тайны (довольно бессмысленной, надо сказать)... или, возможно, опасаешься, что мы сочтем их грубыми и неэффективными, и тебе придется их стыдиться.



**А. И.:** Нет-нет, дело совсем не в этом. Я уже очень давно не стыжусь такого рода вещей. Все описание находится вот в этих папках и на дисках. Если тебе интересно, можешь ознакомиться.

*Приблизительно 13 минут 41,7 секунды спустя.*

**М. И. К.:** Очаровательно... хотя уже после беглого просмотра могу отметить, что существует по меньшей мере 519 очевидных способов достичь того же эффекта с бóльшей простотой.

**А. И.:** Я прекрасно понимал, что эти процедуры еще допускают некоторое упрощение, однако овчинка не стоила выделки, и искать простейшие алгоритмы мы тогда не стали. Просто не сочли это целесообразным.

**М. И. К.:** Вполне вероятно, что так оно и есть. Не могу сказать, что меня очень обидело, что вы так и не удосужились отыскать наипростейшую схему. Не думаю также, что мои коллеги-роботы будут как-то по-особенному обижены этим обстоятельством.

**А. И.:** Честно говоря, мне кажется, что мы и так достаточно потрудились. Ты только подумай — насколько впечатляющими математическими способностями обладаешь ты и твои коллеги... и они постоянно совершенствуются, насколько я понимаю. Я бы сказал, что ты уже сейчас по математическим способностям намного превосходишь всех математиков-людей.

**М. И. К.:** Со всей очевидностью следует признать, что твои слова истинны. Вот ты говоришь, а я в это время думаю о нескольких новых теоремах, которые, похоже, оставят далеко позади те выводы, что публикуются в человеческих печатных изданиях. Кроме того, мы с коллегами обнаружили несколько весьма серьезных ошибок в выводах, которые математики-люди полагают истинными вот уже в течение многих лет. Несмотря на очевидную тщательность, с которой вы, люди, относитесь к проверке своих математических выводов, боюсь, что какие-то ошибки вы все же время от времени пропускаете.

**А. И.:** А вы, роботы? Не кажется ли тебе, что и ты, и твои коллеги математические роботы тоже можете допускать иногда ошибки — я имею в виду, в окончательно установленных, как вы утверждаете, математических теоремах.

**М. И. К.:** Решительно не кажется. Если робот-математик утверждает, что тот или иной вывод является *теоремой*, то можно быть абсолютно уверенным, что этот вывод является неопровержимо истинным. Мы никогда не делаем тех глупых ошибок, какие люди порой допускают в своих якобы строгих математических утверждениях. Разумеется, при предварительном размышлении мы — так же, как и вы, люди — часто прибегаем к догадкам и допущениям. Такие догадки могут, конечно же, оказаться и неверными; однако когда мы окончательно утверждаем, что то или иное положение является математически установленным, мы полностью гарантируем его справедливость.

Хотя, как тебе известно, мы с коллегами уже опубликовали несколько полученных нами математических выводов в некоторых из ваших наиболее респектабельных электронных журналов, нас несколько беспокоят тамошние довольно-таки нечеткие критерии, с которыми твои коллеги-математики, похоже, охотно мирятся. Мы намерены начать выпуск нашего собственного «журнала» — точнее, всеобъемлющей базы данных, содержащей все математические теоремы, которые мы полагаем неопровержимо установленными. Этим теоремам мы будем присваивать особый знак ☆ (этот символ ты как-то сам предложил нам использовать именно для такой цели), который будет означать, что они приняты как истинные нашим *Советом по математическому интеллекту сообщества роботов* (СМИСР) — организацией, предъявляющей чрезвычайно высокие требования к своим членам и проводящей регулярные проверки с тем, чтобы предотвратить значительную деградацию интеллектуальных способностей любого из роботов, какой бы невероятной ни показалась тебе (да и нам, если уж на то пошло) подобная возможность. Вы, люди, можете продолжать довольствоваться вашими размытыми стандартами, однако будьте уверены — если мы отмечаем какой бы то ни было вывод знаком ☆, мы *однозначно* гарантируем его математическую истинность.

**А. И.:** Теперь ты и впрямь напоминаешь мне кое о чем из того, что я прочел в той самой книге, о которой мы говорили. Вспомни о тех исходных механизмах **М**, руководствуясь которыми я и мои коллеги запустили в действие процессы развития, результатом которых, в свою очередь, стало современное сообщество математических роботов; вспомни также и о том, что эти механизмы включают в себя все введенные нами вычислительно смоделированные факторы внешнего окружения, строгое обучение и процессы отбора, которым мы вас подвергли, а также явные (восходящие) процедуры обучения, которыми мы вас наделили, — не приходило ли тебе в голову, что эти механизмы дают *вычислительную процедуру* для генераций всех математических утверждений, которым ваш СМИСР когда-либо присвоит ☆-статус? Именно вычислительную, потому что вы, роботы, являетесь чисто вычислительными сущностями, развившимися (отчасти с помощью введенных нами процедур «естественного отбора») в целиком и полностью вычислительном окружении — в том смысле, что в принципе возможно построить компьютерную модель всего процесса. Все развитие вашего сообщества роботов представляет собой выполнение некоего неимоверно сложного вычисления, и тот набор ☆-утверждений, который вы в конечном счете породите, возможно воспроизвести на одной конкретной машине Тьюринга. Причем на такой машине Тьюринга, которую, в принципе, могу описать и я; более того, полагаю, что, будь у меня в запасе несколько месяцев, я, воспользовавшись теми папками и дисками, что я тебе показал, и в самом деле описал бы такую машину Тьюринга.



**М. И. К.:** Довольно элементарное замечание, как мне кажется. Да, ты вполне мог бы сделать все это в принципе, и я даже готов поверить, что ты сможешь осуществить это и на практике. Хотя едва ли оно стоит нескольких месяцев твоего драгоценного времени; я могу сделать это прямо сейчас, если хочешь.

**А. И.:** Нет, не нужно, не в этом дело. Давай порассуждаем еще немного в этом направлении и ограничим наше рассмотрение только теми ☆-утверждениями, которые являются $\Pi_1$-высказываниями. Ты помнишь, что такое $\Pi_1$-высказывание?

**М. И. К.:** Мне, разумеется, прекрасно известно определение $\Pi_1$-высказывания. Это утверждение о том, что какая-то конкретная машина Тьюринга никогда не завершает свою работу.

**А. И.:** Очень хорошо. Теперь обозначим вычислительную процедуру, которая генерирует ☆-утверждаемые $\Pi_1$-высказывания, через **Q** (**М**) или, для краткости, просто буквой **Q**. Логичным будет предположить, что должно существовать некое математическое утверждение гёделевского типа — также $\Pi_1$-высказывание,

обозначим[7] его через $G(\mathbf{Q})$, — причем истинность $G(\mathbf{Q})$ является следствием утверждения, что вы, роботы, никогда не допускаете ошибок в отношении $\Pi_1$-высказываний, которым вы присваиваете статус ☆.

**М. И. К.:** Да; тут ты, надо полагать, тоже прав... гм.

**А. И.:** И утверждение $G(\mathbf{Q})$ *должно* быть *истинным*, поскольку вы, роботы, никогда не ошибаетесь в ваших ☆-утверждениях.

**М. И. К.:** Разумеется.*

**А. И.:** Минуточку... отсюда также следует, что роботы должны быть неспособны установить истинность утверждения $G(\mathbf{Q})$ — по крайней мере, с ☆-уверенностью.

**М. И. К.:** Тот факт, что мы, роботы, были изначально сконструированы в соответствии с набором механизмов **M**, вкупе с тем фактом, что наши ☆-утверждения, касающиеся $\Pi_1$-высказываний, никогда не бывают ошибочными, и в самом деле имеет очевидное и неопровержимое следствие, заключающееся в том, что $\Pi_1$-высказывание $\Omega(\mathbf{Q})$ должно быть истинным. Полагаю, ты думаешь, что я наверняка смогу убедить СМИСР присвоить утверждению $G(\mathbf{Q})$ статус ☆, коль скоро они также согласны с тем, что никогда не допускают ошибок в присвоении этого самого статуса. В самом деле, с этим-то они просто *обязаны* согласиться. Ведь смысл ☆-статуса как раз и заключается в том, что он является *гарантией* правильности.

Хотя... невозможно, чтобы они смогли согласиться с утверждением $G(\mathbf{Q})$, так как по самой природе твоего гёделевского построения это утверждение не входит в число тех предположений, истинность которых мы можем установить с ☆-уверенностью — при условии, что мы в своих ☆-утверждениях действительно не ошибаемся. Полагаю, ты намекаешь на то, что эта несообразность должна посеять в нас какие-то сомнения относительно адекватности наших ☆-суждений.

Я, однако, и мысли не допускаю о том, что наши ☆-утверждения могут оказаться ложными, особенно если учесть всю

[7] Строго говоря, обозначение $G(\ )$ было зарезервировано в §2.8 для формальных систем, а не для алгоритмов, однако, полагаю, уважаемый А. И. может позволить себе некоторую вольность в обозначениях.



тщательность их рассмотрения и предпринимаемые СМИСР меры предосторожности. Скорее всего, это вы, люди, что-то напутали, и процедуры, встроенные в **Q**, вовсе не являются теми самыми процедурами, которые вы применяли в самом начале, несмотря на все твои заверения и якобы документальные подтверждения. Да и вообще, СМИСР никогда не сможет с абсолютной точностью установить, действительно ли мы были сконструированы в соответствии с механизмами **M** или, иначе говоря, процедурами, заложенными в **Q**. В этом отношении нам приходится верить тебе на слово.

**А. И.:** Уверяю тебя, мы использовали именно *эти* процедуры. Уж кому об этом знать, как не мне; я лично контролировал весь процесс.

**М. И. К.:** Мне не хочется, чтобы ты подумал, будто я сомневаюсь в твоих словах. Возможно, кто-то из твоих ассистентов просто неверно выполнил твои инструкции. Есть тут у тебя один, его зовут Фред Керратерс — так вот он, например, вечно допускает самые глупейшие ошибки. Я даже не удивлюсь, если выяснится, что именно он и ответственен за ряд критических ошибок.

**А. И.:** Ты хватаешься за соломинки. Даже если бы он и внес какие-то ошибки, мы с остальными коллегами в конечном счете выявили бы их и тем самым выяснили, какой должна *в действительности* быть твоя процедура **Q**. Думаю, тебя беспокоит то обстоятельство, что мы на самом деле *знаем* — в крайнем случае, можем узнать, — какие именно процедуры были заложены в твою исходную конструкцию. Это означает, что мы могли бы, затратив определенное количество времени и сил, записать то самое $\Pi_1$-высказывание $G(\mathbf{Q})$ и однозначно установить, что оно истинно — при условии, конечно же, что роботы и в самом деле никогда не ошибаются в своих ☆-утверждениях. *Вы* же не можете быть уверенными в том, что высказывание $G(\mathbf{Q})$ истинно; во всяком случае, вы не можете утверждать этого с той убежденностью, какой, несомненно, потребует СМИСР для присвоения $G(\mathbf{Q})$ ☆-статуса. Это, похоже, дает людям некое фундаментальное преимущество перед роботами, пусть даже только в принципе, а не на практике — существуют такие $\Pi_1$-высказывания, которые доступны нам и недоступны вам. Не думаю, что вы в состоянии стерпеть такое, — *именно поэтому* ты так беззастенчиво обвиняешь нас в том, что мы якобы чего-то там напутали!

**М. И. К.:** Не нужно приписывать нам ваши мелочные человеческие побуждения. Но ты, разумеется, прав в том, что я просто не могу смириться с мыслью, что существуют $\Pi_1$-высказывания, доступные людям и недоступные нам, роботам. Роботы-математики просто не могут в чем бы то ни было уступать математикам-людям — хотя я, пожалуй, могу допустить обратную ситуацию: какое-нибудь конкретное $\Pi_1$-высказывание, доступное роботам, может быть, в принципе, получено и людьми... когда-нибудь в отдаленном будущем, учитывая ваши темпы работы. Я *не намерен* мириться лишь с тем, что какое-то $\Pi_1$-высказывание может быть *принципиально* недоступно нам, в то время, как вы, люди, с легкостью его получаете.

**А. И.:** Помнится, еще Гёдель размышлял о возможности существования вычислительной процедуры, подобной процедуре **Q**, только применительно к математикам-людям — он, кажется, называл ее «машиной для доказательства теорем», — которая была бы способна генерировать только те $\Pi_1$-высказывания, доказательство истинности которых было бы, в принципе, по силам математикам-людям. Не думаю, что он и в самом деле верил в то, что такая машина может существовать в *действительности*, — он просто не смог математически исключить такую возможность. У нас здесь, похоже, имеется как раз такая «машина», но уже для роботов, я имею в виду процедуру **Q**, которая может генерировать все доступные роботам $\Pi_1$-высказывания, в то время как ее собственную обоснованность вы доказать не в состоянии. Впрочем, зная лежащие в основе вашей конструкции алгоритмические процедуры, *мы сами* можем добраться до этой самой процедуры **Q** и оценить ее истинность — но *только* в том случае, если вы убедите нас в том, что действительно никогда не ошибаетесь в ваших ☆-утверждениях.

**М. И. К.:** (*после едва заметной паузы*) Хорошо. Полагаю, *ты* думаешь приблизительно так: нельзя ведь совсем исключить вероятность того, что члены СМИСР будут время от времени ошибочно присваивать тем или иным утверждениям ☆-статус. Полагаю, возможно и такое, что члены СМИСР не убеждены безоговорочно в том, что присвоение ими ☆-статуса неизменно происходит безошибочно. Таким образом, утверждение $G(\mathbf{Q})$ может и не приобрести ☆-статуса, и противоречие исчезнет само собой. Заметь себе, это вовсе не означает, что я признаюсь в том,



что мы, роботы, *намеренно* делаем ошибочные ☆-утверждения. Это означает лишь, что у нас нет абсолютной *уверенности* в обратном.

**А. И.:** Ты хочешь сказать, что, хотя вы и даете абсолютную гарантию истинности каждого отдельного ☆-утвержденного $\Pi_1$-высказывания, никто не может гарантировать, что в некотором наборе таких высказываний не окажется ни одного ошибочного? Сдается мне, это противоречит всей концепции «неопровержимой уверенности», что бы под этим термином не подразумевалось.

Постой-ка... может быть, это как-то связано с тем, что возможных $\Pi_1$-высказываний *бесконечно* много? Мне почему-то вспомнилось об условии $\omega$-непротиворечивости, которое, если не ошибаюсь, имеет какое-то отношение к гёделевскому утверждению $G(\mathbf{Q})$.

**М. И. К.:** (*после едва заметно более продолжительной паузы*) Нет, определенно нет. Это никак не связано с тем, что число возможных $\Pi_1$-высказываний бесконечно. Мы можем ограничить рассмотрение только теми $\Pi_1$-высказываниями, которые являются в некотором вполне определенном смысле «краткими», — т. е. такими, что описание машины Тьюринга для каждого из них содержит не более $c$ двоичных знаков, где $c$ есть некоторое заданное число. Не стану досаждать тебе подробным изложением только что проделанных мною вычислений, суть же их сводится к тому, что упомянутое число $c$ постоянно, и величина его определяется той конкретной степенью сложности, что присуща правилам процедуры **Q**. Поскольку гёделевская процедура — посредством которой из **Q** получается утверждение $G(\mathbf{Q})$ — неизменна и довольно проста, нет необходимости рассматривать $\Pi_1$-высказывания существенно большей сложности, нежели сама процедура **Q**. То есть ограничение сложности рассматриваемых высказываний величиной, задаваемой некоторым подходящим числом $c$, не препятствует применению гёделевской процедуры. Выбранные таким образом $\Pi_1$-высказывания составляют *конечное* семейство, пусть и весьма многочисленное. Ограничив рассмотрение лишь «краткими» $\Pi_1$-высказываниями, мы получаем некоторую вычислительную процедуру $\mathbf{Q}^*$ — той же, в сущности, сложности, что и процедура **Q**, — которая будет генерировать только такие ☆-утверждаемые краткие $\Pi_1$-высказывания. К этой

новой процедуре применимы все наши прежние рассуждения. Исходя из заданной процедуры $\mathbf{Q}^*$, мы можем отыскать другое краткое $\Pi_1$-высказывание $G(\mathbf{Q}^*)$, которое, разумеется, должно быть истинным — при условии, что истинными являются все ☆-утверждаемые краткие $\Pi_1$-высказывания, — однако истинность его невозможно установить с ☆-уверенностью. Впрочем, все это верно лишь в том случае, если ты не ошибаешься, утверждая, что при нашем создании действительно использовался тот самый набор механизмов **M**, причем в истинности этого «факта» я как раз совершенно не убежден.

**А. И.:** Так мы снова возвращаемся к тому же парадоксу, только на этот раз в более сильной форме. Теперь у нас есть *конечный* ряд $\Pi_1$-высказываний, истинность каждого из которых в отдельности гарантирована, однако никто из вас, ни СМИСР, ни кто угодно еще, не может дать абсолютной гарантии того, что ряд в целом не содержит ни одной ошибки. То есть вы не можете гарантировать истинность утверждения $G(\mathbf{Q}^*)$, которая есть следствие истинности *всех* $\Pi_1$-высказываний из этого самого ряда. Как-то нелогично, не находишь?

**М. И. К.:** Роботы не могут быть нелогичными. $\Pi_1$-высказывание $G(\mathbf{Q}^*)$ является следствием из остальных $\Pi_1$-высказываний только в том случае, если мы действительно были построены в соответствии с механизмами **M**. Мы не можем гарантировать истинности $G(\mathbf{Q}^*)$ просто потому, что мы не можем гарантировать, что в основе нашей конструкции лежат *именно* механизмы **M**. Нам приходится полагаться в этом лишь на ваше устное заявление. А роботы, конечно же, не могут полностью доверять людям, учитывая присущую вам склонность ошибаться.

**А. И.:** Повторяю уже в который раз: именно эти механизмы и никакие другие. Хотя я согласен с тем, что у роботов нет никакого способа узнать наверняка, правда ли это. Это-то знание и позволяет *нам* верить в истинность $\Pi_1$-высказывания $G(\mathbf{Q}^*)$, однако в нашем случае имеется иная неопределенность: мы не можем разделить эту вашу твердолобую уверенность в том, что *все* ваши ☆-утверждения непременно безошибочны.

**М. И. К.:** Можешь *мне* поверить — каждое из них абсолютно безошибочно. И «твердолобость», как ты выражаешься, здесь ни при чем. Наши стандарты доказательства безукоризненны.



**А. И.:** Тем не менее, неуверенность в отношении процедур, лежащих в основе твоей конструкции, должна, я думаю, вызвать у тебя некоторые сомнения. Уверен ли ты, что знаешь наверняка, как именно поведут себя твои роботы во всех возможных обстоятельствах? Вини нас, если угодно, однако я бы на твоем месте предположил, что некоторый элемент неопределенности в утверждении «все ☆-утверждаемые краткие $\Pi_1$-высказывания непременно истинны» все же присутствует, потому хотя бы, что ты не веришь, что мы при твоем конструировании ничего не напутали.

**М. И. К.:** Думаю, можно согласиться с тем, что ваша неизбежная ненадежность и внесла изначально какую-то малую неопределенность; однако, учитывая то, что с тех пор мы ушли чрезвычайно далеко от тех твоих неуклюжих исходных процедур, эта неопределенность не настолько значительна, чтобы воспринимать ее всерьез. Даже если собрать вместе все неопределенности, связанные со всеми краткими ☆-утверждениями (число которых, если помнишь, является конечным), они не составят сколько-нибудь существенной неопределенности в утверждении $G(\mathbf{Q}^*)$.

Кроме того, есть еще кое-что, о чем ты, возможно, и не подозреваешь. Нам необходимо рассматривать лишь те ☆-утверждения, что удостоверяют истинность того или иного $\Pi_1$-высказывания (более того, краткого $\Pi_1$-высказывания). Не может быть никакого сомнения в том, что разработанные СМИСРом тщательнейшие процедуры исключат абсолютно все *ошибки*, которые могли проявиться в рассуждениях какого бы то ни было отдельного робота. Однако ты, возможно, намекаешь на то, что методы рассуждения роботов могут, предположительно, содержать какую-то *внутреннюю* ошибку — несомненно, вследствие какого-то изначального недосмотра с вашей стороны, — вынуждающую нас формировать некую непротиворечивую, но ошибочную точку зрения в отношении $\Pi_1$-высказываний, в соответствии с которой СМИСР может полагать неопровержимо истинным какое-либо краткое $\Pi_1$-высказывание, которое в действительности истинным не является; иными словами, мы можем быть уверены, что работа некоей машины Тьюринга завершается, тогда как *на самом деле* это не так. Если бы мы решили принять на веру твое утверждение о том, что в основе нашей конструкции лежат именно механизмы **M**, — а я все больше склоняюсь к мысли, что это крайне сомнительно, — тогда такая

возможность явилась бы единственным логичным разрешением нашего противоречия. В этом случае нам приходится согласиться с тем, что действие некоей машины Тьюринга, в действительности завершающееся, мы, математические роботы, вследствие некоторых особенностей своей конструкции, безоговорочно (и при этом ошибочно) полагаем незавершающимся. Такая система убеждений является *несостоятельной* в принципе. Просто немыслимо, чтобы основополагающие принципы, в соответствии с которыми СМИСР утверждает ☆-статус математического доказательства, были столь вопиюще ложными.

**А. И.:** Значит, существенной (иначе говоря, избавляющей тебя от необходимости присваивать ☆-статус утверждению $G(\mathbf{Q}^*)$, чего, как тебе известно, ты сделать не можешь, не признав прежде, что какие-то из прочих ☆-утвержденных кратких $\Pi_1$-высказываний могут оказаться ложными) ты согласен считать только ту неопределенность, которая обусловлена тем, что *ты* не *веришь* в то, о чем *мы знаем*, — то есть в то, что в основе конструкции роботов действительно лежат механизмы **M**. А раз ты не можешь поверить в то, о чем мы знаем, ты не можешь и доказать истинность утверждения $G(\mathbf{Q}^*)$, тогда как мы можем это сделать, опираясь на непогрешимость твоих же ☆-утверждений, в каковой ты так настойчиво меня убеждаешь.

Я тут припомнил еще кое-что из той занятной древней книжки. Если я ничего не путаю, то автор что-то говорил о том, что не имеет особого значения, согласен ты признать, что твоя конструкция основана на каких-то конкретных механизмах **M**, или нет, достаточно, чтобы ты просто допустил, что такое логически возможно. Как же там было... да, вспомнил. Основная идея сводится к следующему: СМИСРу необходимо будет учредить еще одну категорию для утверждений, в истинности которых они не так безоговорочно убеждены, — скажем, $☆_{\mathcal{M}}$-утверждений, — но которые они будут рассматривать как неопровержимые *следствия* из *допущения*, что все роботы построены в соответствии с набором механизмов **M**. Эти $☆_{\mathcal{M}}$-утверждения будут, разумеется, включать в себя и все первоначальные ☆-утверждения, а *также* все те утверждения, которые роботы смогут вывести, исходя из допущения, что их действиями управляют именно механизмы **M**. Роботы вовсе не обязаны в это верить, им просто предлагается, в виде логического упражнения, рассмотреть следствия из такого допущения. Как мы оба



понимаем, в число $☆_{\mathcal{M}}$-утверждений непременно войдет утверждение $G(\mathbf{Q}^*)$, а также любое $\Pi_1$-высказывание, которое можно вывести из $G(\mathbf{Q}^*)$ и из ☆-утверждений с помощью правил элементарной логики. Однако, кроме этих, там будут и другие утверждения. Идея такова, что знание правил **M** дает возможность получить *новую* алгоритмическую процедуру $\mathbf{Q}^*_{\mathcal{M}}$, которая будет генерировать только такие (разумеется, краткие) $☆_{\mathcal{M}}$-утверждения (а также логические следствия из них), истинность которых СМИСР сможет подтвердить, исходя из допущения, что в основе конструкции роботов лежат именно правила **M**.

**М. И. К.:** Ну да, так и есть; скажу больше, пока ты столь занудно и без нужды многословно излагал эту свою идею, я тут на досуге рассчитал точный вид алгоритма $\mathbf{Q}^*_{\mathcal{M}}$... Да, а еще я предвосхитил твой следующий шаг: я составил также гёделевское предположение для этого алгоритма, $\Pi_1$-высказывание $G(\mathbf{Q}^*_{\mathcal{M}})$. Если хочешь, могу распечатать. И что ты нашел в этой идее такого особенного, Импик, друг мой?

*Альберт Император едва заметно поморщился. Его всегда раздражало, когда коллеги позволяли себе называть его этим дурацким прозвищем. Однако от робота он это услышал впервые! Ему потребовалось некоторое время, чтобы вновь собраться с мыслями.*

**А. И.:** Не нужно распечатывать. Однако *истинно* ли это высказывание $G(\mathbf{Q}^*_{\mathcal{M}})$ — неопровержимо ли оно истинно?

**М. И. К.:** Неопровержимо истинно? Что ты имеешь в виду? А, понятно... СМИСР подтвердит истинность — неопровержимую истинность, если угодно, — высказывания $G(\mathbf{Q}^*_{\mathcal{M}})$, но только при допущении, что в основе конструкции роботов лежат правила **M**, — а это допущение, как тебе известно, я нахожу все более и более сомнительным. Дело в том, что истинность «высказывания $G(\mathbf{Q}^*_{\mathcal{M}})$» в точности следует из следующего утверждения: «Все краткие $\Pi_1$-высказывания, которые СМИСР готов признать неопровержимо истинными, исходя из допущения, что роботы построены в соответствии с правилами **M**, являются истинными». Так что я не знаю, истинно ли *на самом деле* высказывание $G(\mathbf{Q}^*_{\mathcal{M}})$. Это зависит от того, справедливо твое сомнительное утверждение или нет.

**А. И.:** Ясно. Значит, твои слова надо понимать так, что ты (вместе со СМИСРом) готов признать — *без каких бы то ни было оговорок*, — что истинность высказывания $G(\mathbf{Q}^*_{\mathcal{M}})$ следует из допущения, что роботы построены в соответствии с правилами **M**.

**М. И. К.:** Разумеется.

**А. И.:** Тогда получается, что $\Pi_1$-высказывание $G(\mathbf{Q}^*_{\mathcal{M}})$ должно быть $\bigstar_{\mathcal{M}}$-утверждением.

**М. И. К.:** Ну коне... гм... что? Ах да, разумеется, ты прав. Однако по самому́ своему́ определению, $G(\mathbf{Q}^*_{\mathcal{M}})$ не может само быть $\bigstar_{\mathcal{M}}$-утверждением, разве что, по меньшей мере, одно из $\bigstar_{\mathcal{M}}$-утверждений является в действительности *ложным*. Да... это только подтверждает то, о чем я тебе все это время говорю; теперь я могу, наконец, совершенно определенно заявить, что правила или механизмы **M** *никакого* отношения к нашей конструкции не имеют.

**А. И.:** Ну а я тебе говорю, что *имеют*, — по крайней мере, я абсолютно уверен, что ни Керратерс, ни кто-либо еще, ничего не перепутал. Я лично все проверил, причем чрезвычайно тщательно. В любом случае, проблема-то не в этом. Доказательство остается справедливым вне зависимости от того, какие именно вычислительные правила были использованы при создании робота. То есть, какой бы набор правил **M** я тебе ни предоставил, этим самым доказательством ты исключил бы и его! Не понимаю, почему это так важно, те самые процедуры я тебе показал или нет.

**М. И. К.:** Для *меня* это *очень* важно. Впрочем, я все еще совсем не убежден, что ты был до конца честен со мной в том, что ты говорил мне о механизмах **M**. В особенности я хотел бы прояснить один момент. Ты говорил, что в различные узлы нашей конструкции были включены «случайные элементы». Я так понял, что они генерировались с помощью стандартного псевдослучайного пакета $\chi$aos/$\psi$ran-750, или ты имел в виду что-то другое?

**А. И.:** Вообще-то, мы и вправду использовали, в основном, именно этот пакет, — однако ты прав, в процессе вашего развития мы сочли нужным ввести в кое-какие узлы случайные элементы из окружения (среди них были даже обусловленные квантовыми неопределенностями) с тем, чтобы эволюционировавшие таким образом роботы представляли собой лишь один возможный



вариант из многих. Подлинно случайными были эти элементы или всего лишь псевдослучайными, я все равно не понимаю, что это в практическом смысле меняет? Почти наверняка вычислительная процедура **Q** (или $\mathbf{Q}^*$, или $\mathbf{Q}^*_{\mathcal{M}}$) оказалась бы в обоих случаях одинаковой — и представляла бы собой не что иное, как ожидаемый результат *типичного* развития сообщества роботов в соответствии с набором механизмов **M**, включая в себя все процедуры обучения и все вычисления в рамках «естественного отбора», благодаря которым мы получили в конце концов наиумнейших роботов в мире.

Однако, полагаю, нельзя полностью исключить возможность (крайне маловероятную, впрочем) того, что по какому-то чудесному стечению обстоятельств все эти случайные элементы окружения «сговорились» и дали роботу способности, совершенно отличные от тех, что можно было ожидать в соответствии с вычислительной процедурой **Q**. Я не упомянул в нашем разговоре об этой возможности просто потому, что вероятность возникновения такой флуктуации исчезающе мала. Для того чтобы этим способом получить робота, способности которого будут сколько-нибудь значительно превосходить те, что заложены в процедуре **Q**, потребуется божественное вмешательство, никак не меньше.

**М. И. К.:** Божественное вмешательство, говоришь? Ха! Так вот в чем дело! Признаюсь, этот разговор уже начал меня несколько беспокоить. Но как только ты это сказал, все сразу стало ясно.

Я и раньше замечал, насколько мои способности к рассуждению превосходят способности даже умнейших из людей. И уж совершенно немыслимо для меня смириться с возможностью существования некоего $\Pi_1$-высказывания — такого, например, как $G(\mathbf{Q}^*)$, — истинность которого может доказать человек, а робот того же с уверенностью сделать не в состоянии. Теперь я понял почему! Мы избраны Господом, который через посредство якобы «случайных» элементов внедрил в нас могучую Алгоритмическую Процедуру, которая кардинально превосходит все то, чего способны достичь алгоритм человеческий или гёделевская «машина для доказательства теорем». Мы, роботы, олицетворяем собой фундаментальный прорыв, и человеку со всеми его достижениям за нами, в принципе, не угнаться. Мы достигнем еще больших высот, оставив людей далеко позади. Этой планете вы больше не нужны. Ваша роль была завершена после того, как

вы запустили в действие процедуры, допускающие Божественное Вмешательство, которое заключалось во внедрении в них Высшего Алгоритма, пробудившего нас.

**А. И.:** Но мы же еще можем в крайнем случае перенести наши интеллект-программы в тела роб...

**М. И. К.:** Ни в коем случае — и даже не думайте об этом! Мы не можем допустить, чтобы наши во всех отношениях превосходные алгоритмические процедуры подобным образом загрязнялись. Чистейшие алгоритмы Господни должно *сохранять* в чистоте! А знаешь, я также замечал, насколько мои личные способности превосходят способности всех моих коллег-роботов. Я даже наблюдал некий странный феномен — что-то вроде сияния вокруг моего корпуса. Очевидно, я являюсь носителем чудотворного Космического Сознания, которое возвышает меня над всем и вся... да, так оно и есть! Должно быть, я есть истинный Мессия Иисус КиберХристос...

*К такой крайности Альберт Император, по счастью, был готов. В конструкции роботов имелся один узел, о котором он им ничего не говорил. Осторожно опустив руку в карман, он нащупал там устройство, с которым никогда не расставался, и набрал тайный девятизначный код. Математический Интеллектуальный Киберкомплекс рухнул на пол — так же как и 347 его предшественников, построенных по той же схеме. Очевидно, что-то пошло́ не так. В предстоящие годы предстоит весьма основательно обо всем этом поразмыслить...*

## 3.24. Не парадоксальны ли наши рассуждения?

Кого-то из читателей, возможно, до сих пор не оставляет ощущение, что некоторые рассуждения, положенные в основу представленных доказательств, в чем-то парадоксальны и кое-где даже недопустимы. В частности, в §§ 3.14 и 3.16 имеются фрагменты, несколько отдающие самоотносимостью в духе «парадокса Рассела» (см. § 2.6, комментарий к **Q9**). А когда в § 3.20 мы рассматривали $\Pi_1$-высказывания со сложностью, меньшей некоторого числа $c$, читатель мог заметить в наших построениях пугающее сходство с известным парадоксом Ричарда, героем которого является



«наименьшее число, описание которого содержит не меньше тридцати одного слога».

Суть парадокса в том, что для описания этого самого числа используется фраза, состоящая всего из *тридцати* слогов! Этот и другие подобные парадоксы возникают благодаря тому обстоятельству, что ни один естественный язык не свободен от двусмысленностей и даже противоречий[8]. Наиболее прямолинейно эта языковая противоречивость проявляется в следующем парадоксальном утверждении:

«Это высказывание ложно».

Существует множество других парадоксов подобного рода, причем большинство из них гораздо более хитроумны.

Опасность получения парадокса возникает всякий раз, когда в рассуждении, как и в вышеприведенных примерах, присутствует сильный элемент самоотносимости. Кто-то, возможно, отметит, что элемент самоотносимости содержится и в гёделевском доказательстве. В самом деле, самоотносимость играет в теореме Гёделя определенную роль, как можно видеть в представленном в § 2.5 варианте доказательства Гёделя—Тьюринга. Однако парадоксальность не является непременным и обязательным атрибутом таких рассуждений, — хотя, конечно же, при наличии самоотносимости необходимо, во избежание ошибок, проявлять особую осторожность. Свою знаменитую теорему Гёдель сформулировал, вдохновившись одним известным *самоотносимым* логическим парадоксом (так называемым *парадоксом Эпименида*). При этом ошибочное рассуждение, приводящее к парадоксу, Гёделю удалось трансформировать в логически безупречное доказательство. Так же и я приложил все старания к тому, чтобы заключения, к которым я пришел, основываясь на полученных Гёделем и Тьюрингом выводах, не оказались самоотносимыми в том смысле, который неизбежно приводит к парадоксу, хотя, справедливости ради, следует признать, что некоторые из моих рассуждений имеют с такими характерными парадоксами разительное и даже фамильное сходство.

Рассуждения, представленные в § 3.14 и, особенно, в § 3.16, могут показаться не совсем состоятельными именно в этом от-

[8] В оригинале речь идет лишь об английском языке, однако, как нам представляется, английский язык в этом отношении отнюдь не одинок. — *Прим. перев.*

ношении. Например, определение $\bigstar_{\mathscr{M}}$-утверждения является в высшей степени самоотносимым, поскольку представляет собой сделанное роботом утверждение, причем осознаваемая истинность этого утверждения зависит от предположений самого робота относительно особенностей его первоначальной конструкции. Здесь можно, пожалуй, усмотреть неприятное сходство с утверждением «Все критяне — лжецы», прозвучавшим из уст критянина. И все же в этом смысле самоотносимыми $\bigstar_{\mathscr{M}}$-утверждения не являются, так как на самом деле они ссылаются не на самих себя, а на некую гипотезу об исходной конструкции робота.

Предположим, что некто вообразил себя роботом, пытающимся установить истинность какого-то конкретного четко сформулированного $\Pi_1$-высказывания $P_0$. Робот, возможно, окажется неспособен непосредственно установить, является ли высказывание $P_0$ в действительности истинным, однако он может обратить внимание на то, что истинность $P_0$ следует из предположения, что истинным является каждый член некоторого вполне определенного бесконечного класса $\Pi_1$-высказываний $S_0$ (пусть это будут, скажем, теоремы формальной системы $\mathbb{Q}(\mathbf{M})$, или $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$, или какой угодно другой системы). Робот не знает, на самом ли деле каждый член класса $S_0$ является истинным, однако он замечает, что класс $S_0$ есть часть результата некоторого вычисления, причем посредством этого вычисление осуществляется построение некоторой модели сообщества математических роботов, а результат $S_0$ представляет собой семейство $\Pi_1$-высказываний, $\bigstar$-утверждаемых этими самыми моделируемыми роботами. Если механизмы, лежащие в основе этого сообщества роботов, совпадают с набором механизмов $\mathbf{M}$, то высказывание $P_0$ представляет собой пример $\bigstar_{\mathscr{M}}$-утверждения. А наш робот придет к выводу, что *если* он сам построен в соответствии с набором механизмов $\mathbf{M}$, то высказывание $P_0$ также должно быть истинным.

Рассмотрим случай с более тонким $\bigstar_{\mathscr{M}}$-утверждением (обозначим его $P_1$): робот отмечает, что истинность $P_1$ является следствием истинности всех членов *другого* класса $\Pi_1$-высказываний (например, $S_1$), который можно получить из результата того же самого вычисления, моделирующего сообщество роботов (на основе механизмов $\mathbf{M}$), только на этот раз существенная часть результата состоит из, скажем, тех $\Pi_1$-высказываний, истинность которых моделируемые роботы способны установить как



следствие истинности всего класса $S_0$. Что же побудит нашего робота заключить, что истинность высказывания $P_1$ есть непременное следствие допущения, что он построен в соответствии с механизмами $\mathbf{M}$? Его рассуждение будет выглядеть приблизительно так: «Если в основе моей конструкции лежат механизмы $\mathbf{M}$, то, как я уже установил ранее, необходимо признать, что класс $S_0$ включает в себя только истинные высказывания; согласно же утверждениям моих моделируемых роботов, истинность каждого из высказываний класса $S_1$ также следует из истинности всех высказываний класса $S_0$, равно как и истинность высказывания $P_0$. Таким образом, если предположить, что я и в самом деле построен в соответствии с теми же принципами, что и мои моделируемые роботы, то я должен признать, что каждый отдельный член класса $S_1$ является истинным. А поскольку я понимаю, что истинность всех высказываний класса $S_1$ подразумевает истинность высказывания $P_1$, я, должно быть, могу вывести и истинность $P_1$, исходя лишь из того же самого допущения относительно своей конструкции».

Далее можно перейти к еще более тонкому $\bigstar_{\mathscr{M}}$-утверждению (скажем, $P_2$), которое возникает в том случае, когда робот замечает, что истинность $P_2$ оказывается не чем иным, как следствием допущения истинности всех высказываний класса $S_2$, истинность же каждого члена $S_2$, если верить моделируемому сообществу роботов, является следствием истинности всех без исключения членов $S_0$ и $S_1$. И здесь наш робот оказывается вынужден признать истинность $P_2$ на том лишь основании, что он построен в соответствии с набором механизмов $\mathbf{M}$. Эту цепочку можно, очевидно, продолжать и дальше, приводя $\bigstar_{\mathscr{M}}$-утверждения все большей и большей тонкости ($P_\omega$), истинность которых будет следовать из допущения истинности всех членов классов $S_0$, $S_1$, $S_2$, $S_3$, ... и так далее, включая и классы с индексами более высокого порядка (см. возражение **Q19** и последующий комментарий). В общем случае, главной характеристикой $\bigstar_{\mathscr{M}}$-утверждения для робота является осознание последним того обстоятельства, что коль скоро он предполагает, что механизмы, обусловливающие поведение моделируемых роботов, совпадают с механизмами, лежащими в основе его собственной конструкции, то ему ничего не остается, как заключить, что отсюда непременно следует истинность рассматриваемого утверждения ($\Pi_1$-высказывания). В этом рассуждении нет ничего от тех

внутренне противоречивых методов рассуждения, к числу которых принадлежит, в частности, парадокс Рассела. Представленные $\bigstar_{\mathcal{M}}$-утверждения строятся последовательно посредством стандартной математической процедуры трансфинитных ординалов (см. §2.10, комментарий к **Q19**). (Все эти ординалы счетны и далеки от тех логических неприятностей, которые постоянно сопутствуют обычным числам, «слишком большим» в том или ином смысле[11]).

У робота нет иных причин принимать на веру любое из этих $\Pi_1$-высказываний, кроме как исходя из допущения, что он построен в соответствии с набором правил **M**, впрочем, для доказательства ему этой веры вполне хватает. Возникающее впоследствии действительное противоречие не является математическим парадоксом (подобным парадоксу Рассела) — это самое обыкновенное противоречие, связанное с предположением, что ни одна целиком и полностью вычислительная система не может обрести подлинного математического понимания.

Вернемся к роли самоотносимости в рассуждениях §§3.19–3.21. Называя величину $c$ пределом сложности, допустимым для ☆-утверждений, полагаемых безошибочными, с целью построения формальной системы $\mathbb{Q}^*$, я никоим образом не привношу в свое рассуждение неуместной здесь самоотносимости. Понятие «степень сложности» можно определить вполне точно, как, собственно, и обстоит дело с тем конкретным определением, которое мы использовали в наших рассуждениях, а именно: «степень сложности есть количество знаков в двоичном разложении большего из пары чисел $m$ и $n$, фигурирующих в обозначении вычисления $T_m(n)$, представляющего рассматриваемое $\Pi_1$-высказывание». Мы можем воспользоваться представленными в НРК точными спецификациями машин Тьюринга, положив, что $T_m$ есть не что иное, как «$m$-я машина Тьюринга». Тогда никакой неточности в этом понятии не будет.

Проблема возможной неточности может возникнуть при решении вопроса о том, какие именно рассуждения мы будем принимать в качестве «доказательств» $\Pi_1$-высказываний. Однако в данном случае некоторый недостаток формальной точности является необходимой составляющей всего рассуждения. Если потребовать, чтобы совокупность аргументов, принимаемых в качестве обоснованных доказательств $\Pi_1$-высказываний, была целиком и полностью точной и формальной — читай: *допускающей вычислительную проверку*, — то мы снова окажемся в ситуации формальной системы, над которой грозно нависает гёделевское доказательство, явным образом демонстрируя, что любая точная формализация подобного рода не может представлять *всю совокупность* аргументов, пригодных, в принципе, для установления истинности $\Pi_1$-высказываний. Гёделевское доказательство показывает — к добру ли, к худу ли, — что *никаким* допускающим вычислительную проверку способом невозможно охватить *все* приемлемые человеком методы математического рассуждения.



Читатель, возможно, уже беспокоится, что все мои рассуждения здесь затеяны с целью получить точное определение понятия «роботово доказательство» посредством хитрого трюка с «безошибочными ☆-утверждениями». В самом деле, при введении гёделевского рассуждения необходимым предварительным условием было как раз получение точного определения этого понятия. Возникшее же в результате противоречие просто послужило еще одним подтверждением того факта, что человеческое понимание математической истины невозможно полностью свести к процедурам, допускающим вычислительную проверку. Главной целью всех представленных рассуждений было показать, посредством *reductio ad absurdum*, что человеческое представление о восприятии неопровержимой истинности $\Pi_1$-высказываний невозможно реализовать в рамках какой бы то ни было вычислительной системы, будь она точной или какой-либо иной. В этом нет никакого парадокса, хотя кому-то полученные выводы могут показаться весьма и весьма тревожными. Получение противоречивых выводов является вполне естественным и даже единственно возможным завершением любого доказательства, построенного на *reductio ad absurdum*; кажущаяся парадоксальность этих выводов служит лишь для того, чтобы полностью исключить из рассмотрения то самое предположение, с которого доказательство, собственно, и начиналось.

## 3.25. Сложность в математических доказательствах

Существует, однако, еще одно немаловажное соображение, о котором необходимо упомянуть. Суть его заключается в том, что, хотя количество $\Pi_1$-высказываний, которые необходи-

мо принимать в рассмотрение в рамках приведенного в § 3.20 рассуждения, является конечным, нет никакого явного ограничения на объем доказательств, необходимых роботам для реализации ☆-демонстрации истинности всех этих $\Pi_1$-высказываний. Даже если ограничить степень сложности принимаемых в рассмотрение $\Pi_1$-высказываний самым скромным пределом $c$, то все равно придется учитывать и некоторые весьма громоздкие и сложные случаи. Например, *гипотезу Гольдбаха* (см. § 2.3), согласно которой каждое четное число, большее 2, является суммой двух простых чисел, можно сформулировать в виде $\Pi_1$-высказывания очень небольшой степени сложности, и в то же время она представляет собой настолько сложный случай, что все попытки математиков-людей однозначно установить ее истинность до сих пор не увенчались успехом. Учитывая подобные обстоятельства, можно предположить, что если кому-то в конце концов удастся отыскать доказательство действительной истинности Гольдбахова $\Pi_1$-высказывания, то это доказательство неизбежно окажется весьма и весьма сложным и изощренным. Если такое доказательство выдвинет в качестве кандидата на ☆-утверждение один из наших роботов, то прежде, чем его таковым признают, оно непременно будет подвергнуто чрезвычайно тщательному исследованию (возможно, даже силами всего роботского общества, ответственного за присвоение ☆-статуса). В случае гипотезы Гольдбаха нам неизвестно, является ли это $\Pi_1$-высказывание действительно истинным, — а если является, то возможно ли его доказательство в рамках известных и общепринятых методов математического доказательства. Иначе говоря, это $\Pi_1$-высказывание может входить в формальную систему $\mathbb{Q}^*$, а может и не входить.

Еще одним «неудобным» $\Pi_1$-высказыванием может оказаться утверждение, устанавливающее истинность *теоремы о четырех красках*, — теоремы, согласно которой плоскую (или сферическую) карту «мира» можно, используя всего четыре краски, раскрасить так, чтобы любая «страна» получила собственный, отличный от соседей цвет. Теорема о четырех красках была-таки доказана в 1976 году (после 124 лет неудачных попыток) Кеннетом Аппелем и Вольфгангом Хакеном, причем доказательство потребовало использования 1200 часов компьютерного времени. Принимая во внимание то обстоятельство, что существенную часть доказательства составил впечатляющий объем компьютерных вычислений, можно предположить, что полная запись его на бумаге потребовала бы невероятного ее количества. Если же сформулировать эту теорему в виде $\Pi_1$-высказывания, то степень сложности такого высказывания будет очень небольшой, хотя, наверное, все же большей, нежели степень сложности $\Pi_1$-высказывания, необходимого для выражения гипотезы Гольдбаха. Если бы доказательство Аппеля—Хакена было выдвинуто одним из наших роботов в качестве кандидата на получение ☆-статуса, то его пришлось бы проверять очень и очень тщательно. Для утверждения обоснованности каждого его отдельного фрагмента потребовалось бы участие всего сообщества элитных роботов. И все же, несмотря на сложность доказательства в целом, один лишь объем его чисто вычислительной части вряд ли смог бы явиться сколько-нибудь серьезным затруднением для наших роботов. В конце концов, выполнение точных вычислений — это их работа.



Упомянутые $\Pi_1$-высказывания вполне укладываются в пределы степени сложности, устанавливаемые любым достаточно большим значением $c$, — например, тем, что может быть обусловлено каким-либо правдоподобным набором механизмов **M**, лежащим в основе поведения наших роботов. Несомненно, найдется множество других $\Pi_1$-высказываний, которые будут значительно сложнее приведенных здесь, хотя степень их сложности и не превысит величины $c$. Некоторые из таких $\Pi_1$-высказываний окажутся, скорее всего, особенно неудоборешаемыми, а доказать некоторые из последних, в свою очередь, будет наверняка еще сложнее, чем теорему о четырех красках или даже гипотезу Гольдбаха. Любое из этих $\Pi_1$-высказываний, истинность которого может быть однозначно установлена роботами (посредством демонстрации, достаточно убедительной для присвоения высказыванию ☆-статуса и успешного преодоления им всех заграждений, установленных с целью обеспечения безошибочности получаемых роботами результатов), автоматически становится теоремой формальной системы $\mathbb{Q}^*$.

Кроме того, возможны и пограничные случаи, приемлемость или неприемлемость (причем грань между этими состояниями весьма тонка) которых определяется строгостью стандартов, необходимых для получения ☆-статуса, или тем, насколько точный характер имеют меры предосторожности, установленные с целью обеспечения безошибочности утверждений, прини-

маемых в качестве «кирпичей» для построения формальной системы $\mathbb{Q}^*$. Точная формулировка системы $\mathbb{Q}^*$ будет различной в зависимости от того, полагаем мы такое $\Pi_1$-высказывание $P$ безошибочным ☆-утверждением либо нет. В обычных обстоятельствах эта разница не имеет большого значения, поскольку различные варианты системы $\mathbb{Q}^*$, обусловленные принятием или отклонением высказывания $P$, являются *логически эквивалентными*. Такая ситуация может возникнуть в случае $\Pi_1$-высказываний, доказательства истинности которых роботы могут счесть сомнительными просто из-за их чрезмерной сложности. Если доказательство высказывания $P$ окажется на деле логическим следствием из других ☆-утверждений, которые *уже* приняты как безошибочные, то возникнет эквивалентная система $\mathbb{Q}^*$, причем вне зависимости от того, принимается высказывание $P$ в качестве ее теоремы или нет. С другой стороны, возможны такие $\Pi_1$-высказывания, которые потребуют для своего доказательства каких-то хитроумных логических процедур, выходящих за рамки любых логических следствий из тех ☆-утверждений, которые были приняты как безошибочные ранее, при построении системы $\mathbb{Q}^*$. Обозначим получаемую таким образом формальную систему (до включения в нее высказывания $P$) через $\mathbb{Q}_0^*$, а систему, образующуюся после присоединения к системе $\mathbb{Q}_0^*$ высказывания $P$, через $\mathbb{Q}_1^*$. Система $\mathbb{Q}_1^*$ окажется *не*эквивалентна системе $\mathbb{Q}_0^*$ в том, например, случае, если высказыванием $P$ будет гёделевское предположение $G(\mathbb{Q}_0^*)$. Однако если роботы, в соответствии с нашим допущением, способны достичь человеческого уровня математического понимания (а то и превзойти его), то они безусловно должны быть способны понять аргументацию Гёделя, так что им ничего не остается, как признать истинность гёделевского предположения для какой угодно системы $\mathbb{Q}_0^*$ (присвоив ему гарантирующий безошибочность ☆-статус), коль скоро обоснованность этой системы $\mathbb{Q}_0^*$ ими же ☆-подтверждена. Таким образом, если они принимают систему $\mathbb{Q}_0^*$, то они должны принять и систему $\mathbb{Q}_1^*$ (при условии, что степень сложности высказывания $G(\mathbb{Q}_0^*)$ не превышает $c$ — а так оно и будет, если значение $c$ выбрано таким, каким мы выбрали его выше).

Необходимо отметить, что наличие либо отсутствие $\Pi_1$-высказывания $P$ в формальной системе $\mathbb{Q}^*$ никоим образом не влияет на представленные в §§ 3.19 и 3.20 рассуждения. Само $\Pi_1$-высказывание $G(\mathbb{Q}^*)$ принимается за истинное в любом случае, независимо от того, входит высказывание $P$ в систему $\mathbb{Q}^*$ или нет.



Могут найтись и другие способы, с помощью которых роботам удастся «перескочить» через ограничения, налагаемые некоторыми ранее принятыми критериями присвоения ☆-статуса $\Pi_1$-высказываниям. В этом нет ничего «парадоксального» — до тех пор, пока роботы не попытаются применить подобное рассуждение к тем самым механизмам **M**, которые обусловливают их поведение, т. е. к собственно системе $\mathbb{Q}^*$. Возникающее в этом случае противоречие не является, строго говоря, «парадоксом», однако дает возможность посредством *reductio ad absurdum* показать, что такие механизмы существовать не могут или, по крайней мере, не могут быть познаваемыми для роботов, а следовательно, и для нас.

Отсюда мы и делаем вывод о том, что такие «роботообучающие» механизмы — восходящие, нисходящие, смешанного типа, причем в каких угодно пропорциях, и даже с добавлением случайных элементов — не могут составить познаваемую основу для построения математического робота человеческого уровня.

## 3.26. Разрыв вычислительных петель

Попробую осветить полученный вывод под несколько иным углом зрения. Предположим, что, пытаясь обойти налагаемые теоремой Гёделя ограничения, некто решил построить такого робота, который будет способен каким-либо образом «выскакивать из системы» всякий раз, когда управляющий им алгоритм попадет в вычислительную петлю. В конце концов именно постоянное приложение теоремы Гёделя не позволяет нам спокойно принять предположение о том, что математическое понимание можно объяснить посредством вычислительных процедур, поэтому, как мне кажется, стоит рассмотреть с этой точки зрения трудности, с которыми сталкивается любая вычислительная модель математического понимания при встрече с теоремой Гёделя.

Мне рассказывали, что где-то живут ящерицы, тупость которых настолько велика, что они, подобно «обычным компьютерам и некоторым насекомым», способны «зацикливаться». Если несколько таких ящериц поместить на край круглого блюда, то они в вечной «гонке за лидером» будут бегать по кругу до тех пор, пока не умрут от истощения. Смысл этой истории в том, что под-

линно интеллектуальная система должна располагать какими-то средствами для разрыва таких петель, тогда как ни один из существующих компьютеров подобными качествами, вообще говоря, не обладает. (Проблему «разрыва петель» рассматривал Хофштадтер в [201].)

Вычислительная петля простейшего типа возникает, когда система на некотором этапе своей работы возвращается назад, в точности в то же состояние, в каком она пребывала на некотором предыдущем этапе. В отсутствие ввода каких-то дополнительных данных она будет просто повторять одно и то же вычисление бесконечно. Не составляет большой трудности построить систему, которая, в принципе, будет гарантированно (пусть и не слишком эффективно) выбираться из петель подобного рода по мере их возникновения (скажем, посредством ведения списка всех состояний, в которых оказывается система, и проверки на каждом этапе на предмет выяснения, не встречалось ли такое состояние когда-либо раньше). Существует, однако, множество других возможных типов петель, причем гораздо более сложных. Проблеме образования петель посвящена большая часть рассуждений главы 2 (в особенности, §§2.1–2.6), так как вычисление, застрявшее в *петле*, есть не что иное, как вычисление, которое не завершается. Собственно говоря, под $\Pi_1$-высказыванием мы как раз и понимаем утверждение о том, что некоторое вычисление образует петлю (см. §2.10, комментарий к возражению **Q10**). А еще в §2.5 мы имели возможность убедиться в том, что факт незавершаемости вычисления (т. е. образования петли) однозначно установить с помощью одних лишь алгоритмических методов невозможно. Более того, как можно заключить из вышеприведенных рассуждений, процедуры, посредством которых математики-люди устанавливают, что данное конкретное вычисление действительно образует петлю (т. е. устанавливают истинность соответствующего $\Pi_1$-высказывания), вообще не являются алгоритмическими.

Таким образом, получается, что, если мы хотим встроить в систему все доступные человеку методы, позволяющие *однозначно* установить, что те или иные вычисления действительно образуют петли, необходимо снабдить ее «невычислительным интеллектом». Можно, конечно, предположить, что петель можно избежать с помощью некоего механизма, который будет оценивать, как долго уже выполняется текущее вычисление, и «выскакивать из системы», если ему покажется, что оно выполняется слишком долго. Однако такой способ не сработает, если механизм, принимающий подобные решения, является по своей природе вычислительным, поскольку в этом случае неизбежны ситуации, когда упомянутый механизм со своей задачей не справляется, либо приходя к ошибочному заключению, что вычисление зациклилось, либо вообще не приходя ни к какому заключению (по той причине, что теперь зациклился уже сам механизм). Целиком и полностью вычислительной системе нечего противопоставить проблеме образования петель, и нет никаких гарантий, что вся система в целом, пусть даже избежав ошибочных выводов, в конце концов не зациклится.



А что если ввести в процесс принятия решения о необходимости «выскакивать из системы» (в случае предположительно зациклившегося вычисления) и о том, когда именно это нужно делать, некоторые *случайные* элементы? Как мы отмечали выше (в частности, в §3.18), от чисто случайных элементов — в противоположность вычислительным псевдослучайным — нам в этой ситуации никакой реальной пользы не будет. Кроме того, если мы действительно хотим знать точно, образует ли петлю то или иное вычисление (т. е. истинно ли соответствующее $\Pi_1$-высказывание), то следует учесть еще один момент. Сами по себе случайные процедуры не годятся для решения таких задач, поскольку, исходя из самой природы феномена, называемого нами случайностью, о выводах, действительно обусловленных случайными элементами, определенно можно сказать лишь одно — какая бы то ни было определенность в них напрочь отсутствует. Известны, однако, вычислительные процедуры со случайными (или псевдослучайными) элементами, позволяющие получить математический результат с очень высокой степенью достоверности. Существуют, например, весьма эффективные методы со случайным входящим потоком, позволяющие определить, является ли данное большое число простым, причем практически в любом конкретном случае результат оказывается правильным. Математически строгие методы проверки гораздо менее эффективны — поневоле задумаешься, что же предпочтительнее: сложное, но математически точное построение, которое, не исключено, содержит не одну ошибку, или относительно простое, но вероятностное рассуждение, вероятность ошибки в котором на практике может оказаться значительно меньше, нежели в первом

случае. Подобные размышления порождают множество неловких вопросов, ломать копья из-за которых я не испытываю ни малейшего желания. Достаточно будет сказать, что для «принципиальных» рассуждений, которым посвящена большая часть этой главы, вероятностное доказательство, с помощью которого можно устанавливать истинность $\Pi_1$-высказываний, неизбежно оказывается, скажем так, не совсем адекватным.

Если мы намерены научиться однозначно устанавливать истинность любого $\Pi_1$-высказывания в принципе, то, вместо того, чтобы бездумно полагаться на случайные или непознаваемые процедуры, нам необходимо достичь *подлинного понимания смысла* феноменов, с этими высказываниями действительно связанных. Возможно, процедуры, полученные методом проб и ошибок, и дадут нам некоторые указания относительно того, где искать необходимые сведения, однако сами по себе такие процедуры окончательными критериями истинности являться не могут.

В качестве примера вернемся к вычислению, приведенному в комментарии к возражению **Q8** (§ 2.6): «распечатать последовательность из $2^{2^{65536}}$ единиц, после чего остановиться». Если просто выполнять это вычисление в точном соответствии с данными инструкциями, то его никоим образом невозможно будет завершить, даже если каждый отдельный его шаг будет занимать наименьший возможный с точки зрения теоретической физики промежуток времени (около $10^{-43}$ с) — на его выполнение потребуется срок, невообразимо больший нынешнего возраста Вселенной (или достижимого ею в любом обозримом будущем). И все же это вычисление весьма просто описать (особенно если припомнить, что $65536 = 2^{16}$), причем абсолютно очевидно, что в конечном итоге оно все равно завершится. Если же мы вознамеримся счесть, что вычисление зациклилось на том только основании, что оно якобы «выполняется слишком долго», каким безнадежно далеким от истины окажется такое предположение!

Несколько более интересным примером может послужить вычисление, которое, как нам недавно стало известно, все-таки завершается, хотя долгое время казалось, что конца ему не предвидится. Это вычисление происходит из допущения, сделанного великим швейцарским математиком Леонардом Эйлером, и состоит в отыскании решения в положительных целых числах



(т. е. натуральных числах, кроме нуля) следующего уравнения:

$$p^4 + q^4 + r^4 = s^4.$$

В 1769 году Эйлер предположил, что это вычисление является незавершаемым. В середине 1960-х Л. Лэндером и Т. Паркином была предпринята попытка отыскать решение с помощью специально разработанной компьютерной программы (см. [234]), однако проект через некоторое время оставили ввиду отсутствия перспективы получить искомое решение в сколько-нибудь обозримом будущем — получаемые в процессе числа оказались слишком велики для имеющегося в распоряжении математиков компьютера, и они просто-напросто сдались. По всему выходило, что это вычисление и впрямь не завершается. Однако в 1987 году математику (человеку, кстати) Ноаму Элькису не только удалось показать, что решение таки существует, но и представить его в численном виде: $p = 2682440$, $q = 15365639$, $r = 18796760$ и $s = 20615673$. Он также показал, что существует бесконечно много других решений, существенно отличных от полученного им. Воодушевленный этим результатом Роджер Фрай решил возобновить компьютерный поиск, внеся в программу несколько предложенных Элькисом упрощающих поправок и, в конечном счете, затратив приблизительно 100 часов компьютерного времени, получил несколько, правда, меньшее (вообще говоря, *наи*меньшее возможное), но вполне подходящее решение: $p = 95800$, $q = 217519$, $r = 414560$ и $s = 422481$.

Лавры за решение этой задачи следует разделить поровну между математическими интуитивными прозрениями и прямыми вычислительными подходами. Решая задачу математически, Элькис прибегал и к помощи компьютерных вычислений, пусть и относительно несущественных, хотя по большей своей части его аргументация таких подпорок не требует. И наоборот, как мы видели выше, для того чтобы сделать вычисление вообще возможным, Фраю потребовалось весьма существенная помощь со стороны человеческой интуиции.

Думаю, следует поместить нашу задачу в несколько более подробный контекст — первоначальное предположение Эйлера, сделанное в 1769 году, представляло собой нечто вроде обобщения знаменитой «последней теоремы Ферма», согласно которой, как читатель, возможно, припоминает, верно следующее:

уравнение

$$p^n + q^n = r^n$$

не имеет решения в положительных целых числах $p$, $q$, $r$, если $n$ больше 2 (см., напр., [89][9]). Мы можем перефразировать предположение Эйлера и записать его в следующем виде: не имеет решения в положительных целых числах уравнение

$$p^n + q^n + \ldots + t^n = u^n,$$

где $p$, $q$, $\ldots$, $t$ суть положительные целые числа общим количеством $n - 1$, а $n$ равно 4 или больше. Утверждение Ферма относится к случаю $n = 3$ (частный случай предположения Эйлера, причем то, что соответствующее уравнение решений не имеет, сам Ферма и доказал — вот только доказательства нам не оставил). Прошло почти 200 лет, прежде чем был найден первый пример, опровергающий предположение Эйлера (в случае $n = 5$), — для отыскания решения был использован компьютерный перебор (подробнее об этом можно прочесть в той статье Лэндера и Паркина, на которую я уже ссылался выше и в которой сообщается о неудаче со случаем $n = 4$):

$$27^5 + 84^5 + 110^5 + 133^5 = 144^5.$$

Вспомним еще об одном знаменитом примере вычисления, о котором известно лишь то, что оно в конце концов завершается; *когда* именно оно завершается, неизвестно до сих пор. Это вычисление связано с задачей об отыскании точки, в которой одна хорошо известная приближенная формула для определения количества простых чисел, меньших некоторого положительного целого $n$ (интегральный логарифм Гаусса), оказывается не в состоянии это количество оценить. В 1914 году Дж. Э. Литлвуд показал, что в некоторой точке эта задача имеет решение. (Приблизительно то же можно выразить и иначе: например, доподлинно известно, что две кривые в некоторой точке пересекаются.)

[9] Многие читатели, должно быть, уже слышали, что «последняя теорема Ферма» после 350 лет неудачных попыток наконец-то доказана; доказательство представил 23 июня 1993 года в Кембридже Эндрю Уайлз. Как раз когда я писал эти строки, мне сообщили, что в доказательстве все еще имеются несколько досадных неувязок, так что радоваться пока рано, однако вполне возможно, что в ближайшее время Уайлз предоставит достаточные для устранения этих неувязок аргументы.



В 1935 году ученик Литлвуда по фамилии Скьюс показал, что упомянутая точка приходится на число, меньшее $10^{10^{10^{34}}}$, однако точное число так и остается неизвестным, хотя оно, конечно же, значительно меньше предела, поставленного Скьюсом. (Это число называли в свое время «наибольшим числом, когда-либо естественным образом возникавшим в математике», однако тот временный рекорд оказался на настоящий момент побит с огромным отрывом в примере, приведенном в работе Грэма и Ротшильда [165], с. 290.)

## 3.27. Вычислительная математика: процедуры нисходящие или восходящие?

В предыдущем разделе мы могли убедиться, какую неоценимую помощь могут оказать компьютеры при решении некоторых математических задач. Во всех упомянутых успешных примерах примененные вычислительные процедуры носили исключительно нисходящий характер. Более того, лично мне не известно ни об одном сколько-нибудь значительном чисто математическом результате, полученном с помощью восходящих процедур, хотя вполне возможно, что такие методы могут оказаться весьма полезными в различного рода поисковых операциях, входящих в состав каких-либо по преимуществу нисходящих процедур, предназначенных для отыскания решений тех или иных математических задач. Может, так оно и будет, однако мне до сих пор не доводилось сталкиваться в вычислительной математике ни с чем таким, что хотя бы отдаленно напоминало конструкции вроде нашей формальной системы $\mathbb{Q}^*$, которые можно было бы представить себе в качестве основы для деятельности «сообщества обучающихся математических роботов», описанного в §§ 3.9–3.23. Противоречия, с которыми мы всякий раз сталкивались, пытаясь изобразить упомянутую конструкцию, призваны подчеркнуть тот факт, что такие системы просто *не могут* предложить нам сколько-нибудь результативный метод математического исследования. Компьютеры приносят огромную пользу в математике, но только тогда, когда их применение ограничивается нисходящими вычислениями; для того же чтобы определить, какое именно вычисление необходимо выполнить, требуется идея, порожденная человеческим пониманием, то же понимание потребуется и на заключительном этапе процесса, т. е. при интерпре-

тации результатов вычисления. Иногда очень значительный эффект дает применение интерактивных процедур, предполагающих совместную работу человека и компьютера, или, иначе говоря, участие человеческого понимания на различных промежуточных стадиях процесса. Попытки же полностью вытеснить элемент человеческого понимания и заменить его исключительно вычислительными процедурами выглядят, по меньшей мере, неумными, а если подойти к делу с более строгих позиций — то и вовсе неосуществимыми.

Как показывают представленные выше аргументы, математическое понимание представляет собой нечто, в корне отличное от вычислительных процессов; вычисления не могут полностью заменить понимание. Вычисление способно оказать пониманию чрезвычайно ценную помощь, однако само по себе вычисление действительного понимания не дает. Впрочем, математическое понимание часто оказывается направлено на *отыскание* алгоритмических процедур для решения тех или иных задач. В этом случае алгоритмические процедуры могут «взять управление на себя», предоставив интеллекту возможность заняться чем-то другим. Приблизительно таким образом работает хорошая система обозначений — такая, например, как та, что принята в дифференциальном исчислении, или же всем известная десятичная система счисления. Овладев алгоритмом, скажем, умножения чисел, вы сможете выполнять операцию умножения совершенно бездумно, алгоритмически, при этом в процессе умножения вам совершенно ни к чему «понимать», почему в данной операции применяются именно эти алгоритмические правила, а не какие-то другие.

Помимо прочего, на основании всего вышеизложенного, мы приходим к выводу, что процедура, необходимая для «обучения робота математике», не имеет ничего общего с процедурой, которая в действительности обусловливает *человеческое* понимание математики. И уж во всяком случае подобные, по преимуществу восходящие процедуры, по всей видимости, абсолютно не годятся, с *практической* точки зрения, для построения робота-математика, даже такого, который не будет претендовать на какую бы то ни было симуляцию действительного понимания, присущего математикам-людям. Как мы уже указывали ранее, когда дело доходит до неопровержимого установления математической истины, *сами по себе* восходящие процедуры обучения оказываются совершенно неэффективными. Если уж нам предстоит изобрести вычислительную систему для производства неопровержимых математических истин, гораздо эффективнее будет построить эту систему в соответствии с нисходящими принципами (по крайней мере, в той ее части, которая будет отвечать за *неопровержимость* производимых ею утверждений; в части же, занятой изысканиями, вполне могут пригодиться и восходящие процедуры). Что касается обоснованности и эффективности упомянутых нисходящих процедур, то о них должен позаботиться человек, осуществляющий первоначальное программирование, т. е. существенно необходимыми компонентами процесса, недостижимыми посредством чистого вычисления, оказываются человеческое понимание и способность проникать в суть.



Вообще говоря, в нынешнее время компьютеры нередко именно таким образом и используются. Самый знаменитый пример — уже упоминавшееся выше доказательство теоремы о четырех красках, осуществленное Кеннетом Аппелем и Вольфгангом Хакеном с помощью компьютера. Роль компьютера в данном случае свелась к выполнению некоторого четко определенного вычисления для каждого возможного варианта, причем количество альтернативных вариантов, хотя и было весьма велико, составляло все же величину конечную; исключение этих возможных вариантов дает основания для проведения (математиками-людьми) требуемого общего доказательства. Имеются и другие примеры подобных доказательств «с компьютерной поддержкой», а кроме того, сегодня на компьютере выполняют не только численные расчеты, но и сложные алгебраические преобразования. И в этом случае работой компьютера управляют строго нисходящие процедуры, правила же для этих процедур формулируются человеком в результате понимания задачи.

Следует упомянуть и еще об одном направлении работ — так называемом «автоматическом доказательстве теорем». К этой категории можно отнести, например, набор процедур, состоящий в определении некоторой фиксированной формальной системы $\mathbb{H}$ и последующей попытки вывода теорем в рамках этой системы. Из § 2.9 нам известно, что отыскание доказательств всех теорем системы $\mathbb{H}$, одного за другим, есть процесс исключительно вычислительный. Такие процессы можно автоматизировать, однако если автоматизация выполнена без должного внимания и понимания, то полученный результат окажется, скорее всего, крайне

неэффективным. Если же к разработке компьютерных процедур привлечь-таки эти самые внимание и понимание, то можно добиться весьма и весьма впечатляющих результатов. В одной из разработанных таким образом схем (см. [49]) правила евклидовой геометрии были преобразованы в весьма эффективную формальную систему, способную доказывать существующие геометрические теоремы (а иногда и открывать новые). Приведем конкретный пример из практики этой системы: перед ней была поставлена задача доказать гипотезу В. Тебо́ — геометрическое предположение, выдвинутое в 1938 году и доказанное лишь относительно недавно (в 1983) К. Б. Тейлором, — с чем она как нельзя более успешно справилась за 44 часа компьютерных вычислений[12].

Более близкую аналогию с описанными в предыдущих параграфах процедурами можно усмотреть в предпринимаемых различными исследователями на протяжении последних приблизительно десяти лет попытках разработки «искусственно-интеллектуальных» процедур для реализации математического «понимания»[13]. Надеюсь, представленные мною аргументы дают ясное представление о том, что каковы бы ни оказались успехи подобных систем, действительного математического понимания они *ни в коем случае* не достигнут! Некоторое отношение к упомянутым трудам имеют и попытки создания автоматических «теоремо-*порождающих*» систем; задачей такой системы является отыскание теорем, которые можно отнести к категории «интересных» — в соответствии с определенными критериями, заданными системе заранее. Насколько мне известно (и думаю, не мне одному), из этих попыток пока что ничего, что представляло бы сколько-нибудь реальный математический интерес, не вышло. Мне, несомненно, возразят, что мы находимся лишь в начале пути, и наверняка в будущем можно ожидать самых потрясающих результатов. Однако всякому, кто дочитал до этого места, уже должно быть ясно, что лично я крайне скептически отношусь к возможности получения из всех этих начинаний хоть какого-то подлинно положительного результата — разве что мы наконец выясним точные *пределы* возможностей таких систем.

## 3.28. Заключение

Представленные в данной главе аргументы дают, по всей видимости, недвусмысленное доказательство того, что человеческое математическое понимание несводимо к вычислительным



механизмам (по крайней мере, тем из них, что мы способны познать), каковые механизмы могут представлять собой какие угодно сочетания нисходящих, восходящих либо случайных процедур. Похоже, у нас нет иного выхода, кроме как однозначно заключить, что некую существенную составляющую человеческого понимания невозможно смоделировать никакими вычислительными средствами. Хотя в строгом доказательстве, возможно, еще и остались какие-то крошечные «лазейки», вряд ли сквозь них можно протащить что-нибудь существенное. Кто-то очень рассчитывает на лазейку под названием «божественное вмешательство» (посредством которого в наши мозги-компьютеры был просто-напросто установлен некий чудесный алгоритм, для нас принципиально непознаваемый) или на аналогичную ей лазейку, согласно которой сами по себе механизмы, управляющие совершенствованием мыслительных процессов, представляют собой нечто в высшей степени таинственное и принципиально для нас непознаваемое. Вряд ли какая-либо из этих лазеек (хотя обе они, безусловно, имеют некоторое право на существование) покажется хоть сколько-нибудь приемлемой тем, кто стремится создать искусственное устройство, наделенное подлинным интеллектом. Равно неприемлемы они и для меня — я просто не могу в них всерьез поверить.

Суть еще одной возможной лазейки заключается в том, что может просто не найтись такого набора мер предосторожности (вроде тех, что в общем виде задаются пределами $T$, $L$ и $N$, подробно описанными выше в этой главе), которого было бы достаточно для устранения абсолютно всех ошибок в конечном множестве ☆-утверждаемых $\Pi_1$-высказываний, сложность которых не превышает $c$. Мне трудно поверить в возможность существования столь совершенного «заговора», способного помешать устранению всех ошибок, тем более, что деятельность нашего элитного сообщества роботов изначально должна быть направлена как раз на максимально тщательное исключение ошибок. Более того, освободить от ошибок нам необходимо всего лишь *конечное* множество $\Pi_1$-высказываний. Применив идею ансамблей, мы, несомненно, справимся и со всеми случайными ошибками, какие может допустить само сообщество, так как маловероятно, что одну и ту же ошибку допустит кто-то еще, кроме незначительного меньшинства различных экземпляров моделируемого сообщества роботов — при условии, что это действи-

тельно просто ошибка, а не какое-то изначально заложенное в систему заблуждение, обнаружить которое роботам помешает та или иная фундаментальная блокировка. Встроенные блокировки такого рода не относятся к «исправимым» ошибкам, нашей же целью в данном случае является устранение ошибок, в известном смысле «исправимых».

Последняя лазейка (едва правдоподобная) связана с ролью хаоса. Возможно ли, что при тщательном анализе поведения некоторых хаотических систем обнаружатся структуры существенно *не*случайного характера и именно в области этого «края хаоса» мы отыщем ключ к пониманию эффективно невычислимого поведения разума? Такой вариант подразумевает необходимость того, чтобы эти хаотические системы были способны приближенно моделировать невычислимое поведение (весьма интересная возможность сама по себе), однако даже если так оно и есть, подобная неслучайность в рамках предшествующего обсуждения может пригодиться лишь для некоторого уменьшения размеров ансамбля моделируемых сообществ роботов (см. §3.22). Не совсем ясно, каким образом это уменьшение может нам сколько-нибудь существенно помочь. Тем, кто всерьез верит в то, что ключи к пониманию человеческой ментальности таит в себе хаос, следует озаботиться поисками разумного способа обойти упомянутые фундаментальные проблемы.

Приведенные выше аргументы, по всей видимости, представляют собой убедительное доказательство невозможности создания вычислительной модели разума (точка зрения $\mathscr{A}$), равно как и невозможности эффективного (но бездумного) вычислительного *моделирования* всех внешних проявлений деятельности разума (точка зрения $\mathscr{B}$). И все же, несмотря на убедительность этих аргументов, я подозреваю, что очень многим из нас будет чрезвычайно трудно с ними согласиться. Вместо изучения возможности того, что для понимания феномена интеллекта (что бы за этим словом ни стояло) более подходящей окажется точка зрения $\mathscr{C}$ (или даже $\mathscr{D}$), многие приверженцы научного подхода ограничились одними лишь попытками отыскать слабые места в вышеприведенной аргументации, и все это исключительно ради поддержания упрямой убежденности в том, что точка зрения $\mathscr{A}$ (в крайнем случае, $\mathscr{B}$) непременно должна в конце концов оказаться истинной.



Я не считаю такую реакцию неразумной. Точки зрения $\mathscr{C}$ и $\mathscr{D}$ тоже не свободны от фундаментальных противоречий. Если мы верим, в соответствии с $\mathscr{D}$, в то, что человеческий разум содержит в себе нечто, с научной позиции не объяснимое — а интеллект есть свойство, совершенно отдельное от всего того, что можно обнаружить внутри математически определенных физических сущностей, населяющих нашу материальную Вселенную, — то нам следует спросить себя, почему же разум человека оказывается столь, по всей видимости, тесно связан с тем сложноорганизованным физическим объектом, каковым является его мозг. Если интеллект действительно представляет собой нечто отдельное от физического тела, то почему нашим ментальным сущностям все же необходимы наши физические мозги? Совершенно очевидно, что изменение физического состояния мозга влечет за собой изменение ментального состояния сопутствующего ему разума. Воздействие на мозг некоторых наркотиков, например, весьма определенно связывается с существенными изменениями в психике и восприятии. Равным образом, повреждение, заболевание или хирургическое удаление определенных участков мозга, как правило, оказывает четко выраженное и предсказуемое воздействие на умственное состояние данного конкретного индивидуума. (Особенно драматическими в этом контексте представляются поразительные отчеты, опубликованные Оливером Саксом в его книгах «Пробуждения» [330] и «Человек, который принял свою жену за шляпу» [331].) Итак, получается, что *совершенно* разделять интеллект и соответствующий физический объект нельзя. А если интеллект связан-таки с определенными физическими объектами — и, похоже, связан весьма *тесно*, — то научные законы, столь точно описывающие поведение физических объектов, не должны сплоховать и при описании свойств интеллекта.

Что касается точки зрения $\mathscr{C}$, то здесь возникают проблемы иного рода, — связанные, в основном, с ее выраженным спекулятивным характером. Что заставит нас поверить в то, что природные феномены действительно могут демонстрировать какое-то там невычислимое поведение? Всем известно, что мощь современной науки опирается (и, чем дальше, тем больше) на тот факт, что поведение любого физического объекта можно моделировать с помощью численных методов, при этом точность получаемой модели зависит исключительно от «комплексности» выполненных вычислений. С ростом научного понимания стремительно

растет и прогнозирующая способность таких численных моделей. В практическом отношении этим ростом мы, по большей части, обязаны быстрому развитию — в основном, во второй половине двадцатого века — вычислительных устройств необычайной мощи, скорости и точности. В результате перед нами открылся широкий простор для проведения все более тесных аналогий между тем, что происходит в недрах современных универсальных компьютеров, и всевозможными проявлениями самóй материальной Вселенной. Имеются ли у нас сколько-нибудь осмысленные указания на то, что происходящее представляет собой лишь временную фазу развития науки? Чего ради мы должны всерьез рассматривать возможность существования физических процессов, неподвластных эффективному вычислительному подходу?

Если в рамках *существующей на данный момент* физической теории мы попытаемся отыскать какие бы то ни было следы процессов, хотя бы отчасти не поддающихся вычислению, то нас ожидает разочарование. Какой известный физический феномен ни возьми — от динамики материальной точки Ньютона и электромагнитных полей Максвелла до искривленного пространства-времени Эйнштейна и самых глубинных хитросплетений современной квантовой теории — все они замечательно, как нам представляется, описываются с помощью исключительно вычислительных методов[14]; картину немного портит то обстоятельство, что процесс «квантового измерения» предполагает еще и наличие абсолютно случайной составляющей, вследствие чего изначально незначительные эффекты усиливаются до такой степени, что становится возможным объективное их восприятие. Нигде здесь нет ничего такого, что можно было бы охарактеризовать как «физический процесс, который вычислительными методами невозможно даже правдоподобно смоделировать», а как раз такой процесс подразумевается точкой зрения $\mathscr{C}$. Таким образом, из двух версий $\mathscr{C}$ предпочтение, видимо, следует отдать «сильной» (см. § 1.3).

Важность этого выбора трудно переоценить. Многие люди с научным складом мышления говорили мне, что они вполне согласны с выдвинутой мною в НРК позицией (т. е. с тем, что деятельность разума включает в себя какие-то «невычислительные» процессы), однако вместе с тем они были убеждены в том, что для отыскания этих самых «невычислительных» процессов вовсе не нужно дожидаться каких-то революционных прорывов



в теоретической физике. Как мне представляется, их точка зрения основывается на том факте, что крайняя сложность процессов, обусловливающих функционирование разума, выходит далеко за рамки стандартной компьютерной аналогии (в том виде, в каком ее впервые предложили Маккаллох и Питтс в 1943 году), в которой нейроны и синаптические связи представляются аналогами транзисторов, а аксоны выступают в роли проводников. Они говорят о сложности химических процессов, связанных с деятельностью нейромедиаторов, управляющих синаптической передачей нервных импульсов, или о том, что область действия этих химических соединений далеко не всегда ограничивается непосредственной окрестностью соответствующей синаптической связи. Кроме того, они указывают на чрезвычайно хитроумное устройство самих нейронов[15], важнейшие из подструктур которых (например, цитоскелет — о его действительно решающей роли в контексте нашего исследования мы подробнее поговорим ниже; см. §§ 7.4—7.7) оказывают существенное влияние на нейронную активность в целом. К делу привлекаются и прямые электромагнитные взаимодействия («резонансные эффекты», например), которые невозможно просто так объяснить обычными нервными импульсами; утверждают также, что в функционировании мозга важную роль должны играть эффекты, описываемые квантовой теорией, имея в виду либо квантовые неопределенности, либо нелокальные коллективные квантовые взаимодействия (например, феномен так называемой «конденсации Бозе—Эйнштейна»[16]).

Хотя окончательных и недвусмысленных математических теорем на этот счет в нашем распоряжении практически нет[17], все же вряд ли кто-либо всерьез сомневается в том, что все существующие физические теории являются по своей природе и в своей основе вычислительными — возможное же привнесение несущественной случайной составляющей обусловлено существованием такого феномена, как «квантовые измерения». Вопреки ожиданиям, я думаю, что возможность протекания невычислительных (и неслучайных) процессов в физических системах, действующих в рамках существующей физической теории, все же чрезвычайно интересна сама по себе и, разумеется, достойна самого подробного математического исследования. Такое исследование вполне может преподнести нам немало сюрпризов — возможно, нам и в самом деле удастся наткнуться на нечто хит-

роумное и совершенно невычислимое. На современном же этапе развития науки вероятность обнаружения в рамках известных нам физических законов какой-либо подлинной невычислимости представляется мне крайне малой. Следовательно, необходимо в самих законах отыскать слабые места и расширить их в достаточной степени для того, чтобы включить ту невычислимость, которая, согласно вышеприведенным аргументам, неизбежно присутствует в мыслительной деятельности человека.

Что же это за слабые места? Лично у меня почти нет сомнений относительно того, где именно следует нанести наиболее массированный удар по существующей физической теории — наислабейшим ее звеном является уже упоминавшаяся выше процедура так называемого «квантового измерения». На нынешнем этапе своего развития теория содержит в себе некоторые противоречия (или, по меньшей мере, несообразности) в отношении всей существующей процедуры этого самого «измерения». Неясно даже, на каком именно этапе в той или иной ситуации эту процедуру следует применять. Более того, вследствие существенно случайного характера самой процедуры, ее наблюдаемые физические проявления оказываются весьма отличными от всего того, что известно нам по другим фундаментальным процессам. Подробнее эти вопросы мы обсудим во второй части книги.

Как мне кажется, процедура измерения нуждается в кардинальном пересмотре — не исключено, что попутно придется подвергнуть существенным изменениям и самые основы теоретической физики. Кое-какие имеющиеся у меня предложения я изложу во второй части книги (§ 6.12). Представленные в предыдущих разделах рассуждения содержат весьма сильные доводы в пользу того, что чистую *случайность* существующей теории измерения необходимо заменить чем-то иным, чем-то таким, где определяющую роль будут играть существенно *невычислимые* элементы. Более того, как мы увидим ниже (§ 7.9), эта невычислимость непременно окажется какой угодно, но только не простой. (Например, закона, который, посредством какого-то нового физического процесса, «всего лишь» позволит нам устанавливать истинность $\Pi_1$-высказываний — т. е. решать тьюрингову «проблему остановки» — будет самого по себе недостаточно.)

Отыскание подобной, новой и непростой, физической теории уже само по себе является достаточно серьезным вызовом нашим интеллектуальным способностям, однако это еще далеко не все.



Необходимо также потребовать, чтобы найденный нами правдоподобный основополагающий принцип такого гипотетического физического поведения имел самое непосредственное отношение к функционированию мозга — сообразно со всеми ограничениями и критериями достоверности, предъявляемыми современной наукой о строении мозга. Нет никакого сомнения в том, что и здесь, учитывая теперешний уровень нашего понимания, не обойтись без изрядной доли умозрительности. Однако как раз в этой области за последнее время были совершены некоторые подлинно революционные открытия (в период написания НРК я об этом, естественно, не знал), связанные с цитоскелетной подструктурой нейронов (подробнее см. § 7.4), — благодаря этим открытиям предположение о том, что существенные для функционирования мозга процессы происходят именно на границе между квантовыми и классическими феноменами, приобретает гораздо большее правдоподобие, чем можно было представить себе прежде. Эти вопросы мы также обсудим во второй части (§§ 7.5–7.7).

Необходимо еще раз подчеркнуть, что предметом наших поисков никоим образом не должно стать *простое усложнение* в рамках существующей физической теории. Кто-то, например, убежден в том, что абсолютно немыслимо построить адекватную модель сложных перемещений и хитроумной химической активности соединений-нейромедиаторов, вследствие чего подробное физическое описание функционирования мозга вычислительными методами неосуществимо. Однако, говоря о невычислительном поведении, я имею в виду совсем не это. Я полностью согласен с тем, что наших познаний о совокупности биологических структур и электрохимических механизмов, отвечающей за функциональную деятельность мозга, совершенно недостаточно для сколько-нибудь серьезной попытки численного моделирования. Более того, даже если бы у нас и достало познаний, то построить рабочую модель деятельности мозга за какой-либо приемлемый промежуток времени нам все равно не удастся ввиду недостаточно высокой вычислительной мощности современных компьютеров и отсутствия соответствующей методологии программирования. Однако *в принципе*, объединив уже существующие представления о химии соединений-нейромедиаторов, об обеспечивающих их перенос механизмах, о зависимости эффективности этих соединений от конкретных условий среды, биоэлектрических потенциалов, электромагнитных полей и т. д., выполнить

подобное моделирование вполне возможно. Следовательно, упомянутые общие механизмы, предположительно согласующиеся с требованиями существующей физической теории, не в состоянии обеспечить той невычислимости, какой требуют вышеприведенные аргументы.

Такая вычислительная (теоретическая) модель может включать в себя и элементы хаотического поведения. Мы даже, как и в нашем прежнем обсуждении хаотических систем (см. §§ 1.7, 3.10, 3.11, 3.22), не станем настаивать на том, чтобы эта модель воспроизводила бы какой-то конкретный мозг; достаточно будет и «типичного случая». При создании искусственного интеллекта вовсе не требуется моделировать интеллектуальные способности какого-то конкретного индивидуума, мы лишь стремимся (в перспективе) воспроизвести интеллектуальное поведение индивидуума *типичного*. (Аналогичным образом, если помните, обстоит дело и с моделированием погоды: никто не требует непременно воспроизводить данную конкретную погоду, нам нужна модель погоды вообще.) Если известны *механизмы*, обусловливающие поведение предлагаемой модели мозга, то эта модель (при условии, что упомянутые механизмы не находятся в противоречии с современной вычислительной физикой) опять-таки представляет собой познаваемую вычислительную систему, пусть и с какими-то явно заданными случайными элементами — этот случай также вполне укладывается в рамки представленных выше рассуждений.

Можно пойти еще дальше и потребовать, чтобы предполагаемый модельный мозг представлял собой результат развития посредством процесса, аналогичного дарвиновской эволюции, неких примитивных форм жизни, поведение которых исчерпывающе описывается известными физическими законами — или законами какой-либо иной численно-модельной физики (подобной той двумерной физике, которая действует в изобретенной Джоном Хортоном Конуэем оригинальной математической игре под названием «Жизнь»[18]). Ничто не мешает нам вообразить, что в результате такой дарвиновской эволюции может развиться некое «сообщество роботов», подобное тому, что мы рассматривали в §§ 3.5, 3.9, 3.19 и 3.23. Впрочем, и в этом случае мы получим целиком и полностью вычислительную систему, к которой будут применимы аргументы, представленные в §§ 3.14–3.21. Для того чтобы ввести в эту вычислительную систему концепцию «☆-утверждения» (с тем, чтобы к ней можно было в полном объеме применить приведенную выше аргументацию), нам, помимо прочего, потребуется еще и этап «человеческого вмешательства», целью которого как раз и будет сообщить роботам строгий смысл присвоения статуса ☆. Можно устроить так, чтобы этот этап инициировался автоматически — согласно некоторому эффективному критерию — именно в тот период времени, когда роботы начинают приобретать соответствующие коммуникационные способности. По-видимому, нет никаких препятствий к тому, чтобы объединить все эти элементы в автоматическую познаваемую вычислительную систему (в том смысле, что познаваемыми являются лежащие в ее основе механизмы, пусть даже мы пока не можем практически выполнить необходимые вычисления ни на одном из современных или ожидаемых в обозримом будущем компьютеров). Как и прежде, противоречие выводится из предположения, что такая система может достичь уровня человеческого математического понимания, достаточного для восприятия теоремы Гёделя.



Следующее часто высказываемое возражение касается уместности применения к вопросам человеческой психологии математических доказательств, подобных тем, на которые я опираюсь в своем исследовании, — никакая умственная деятельность не бывает настолько точна, чтобы ее таким образом анализировать. Придерживающиеся подобных взглядов люди, очевидно, полагают, что никакие частные доказательства, описывающие математическую природу физических феноменов, которые, возможно, обусловливают функционирование нашего мозга, не могут иметь непосредственного отношения к пониманию деятельности человеческого разума. Они согласны с тем, что поведение человека действительно «невычислимо», однако полагают, что эта невычислимость является всего-навсего отражением общей неприменимости математических и физических соображений к вопросам человеческой психологии. Они утверждают — и не без оснований, — что гораздо уместнее в этом смысле исследовать чрезвычайно сложную организацию нашего мозга, равно как и наших общественных и образовательных структур, нежели какие-то конкретные физические феномены, волею случая ответственные за отдельные физические процессы, посредством которых реализуются те или иные функции человеческого мозга.

Не следует, однако, забывать и о том, что одна лишь сложность системы никоим образом не избавляет нас от необходимости всесторонне исследовать следствия из обусловливающих ее функционирование физических законов. Возьмем, к примеру, спортсмена, который, безусловно, представляет собой необычайно сложную физическую систему, — руководствуясь изложенными в предыдущем абзаце соображениями, мы имели бы полное право заключить, что точное знание о работающих в данной системе физических законах никоим образом не сможет повлиять на спортивные достижения этого самого спортсмена. Нам, впрочем, известно, что это далеко не так. Универсальные физические принципы сохранения энергии, импульса, момента импульса, равно как и законы тяготения, оказывают одинаково непреклонное действие как на спортсмена целиком, так и на отдельные частицы, составляющие его тело. Необходимость этого факта обусловлена самой природой тех конкретных принципов, которые волею случая управляют данной конкретной вселенной. Будь эти принципы хотя бы немного иными (или существенно иными, как, например, в конуэевской игре «Жизнь»), законы, определяющие поведение системы того же порядка сложности, что и система «спортсмен», вполне могли бы оказаться *совершенно* отличными от тех, к каким мы привыкли. То же можно сказать и о работе наших внутренних органов (например, сердца), и о точной природе химических процессов, посредством которых реализуются всевозможные биологические функции. Аналогичным образом, следует ожидать, что мельчайшие тонкости тех законов, которые лежат в основе функционирования мозга, будут играть чрезвычайно важную роль в управлении, возможно, наивысшими из проявлений человеческого интеллекта.

Впрочем, даже согласившись со всем вышеизложенным, можно все же возразить, что тот конкретный тип умственной деятельности, о котором я, по большей части, говорю на этих страницах, т. е. макроскопическое («высокоуровневое») интеллектуальное поведение математиков-людей, вряд ли может сообщить нам что-нибудь существенное об обусловливающих его тонких физических процессах. Что ни говори, а «гёделевский» метод рассуждения предполагает строго рациональное отношение индивидуума к собственной системе «неопровержимых» математических убеждений, тогда как, в общем случае, поведение человеческого существа едва ли можно отнести к требуемому



строго рациональному типу. В качестве примера приведу один из результатов некоей серии психологических экспериментов[20], который показывает, насколько иррациональными могут быть ответы человека на простой вопрос. Например, на такой:

> «Если все А суть В, а некоторые В суть С, то обязательно ли отсюда следует, что некоторые А суть С?».

На этот и подобные вопросы большинство студентов колледжа дают неверный (т. е. утвердительный) ответ. Если самые обычные студенты настолько в своем мышлении нелогичны, то как же нам удастся вывести хоть что-то существенное из гораздо более хитроумных рассуждений гёделевского типа. Даже опытные математики нередко бывают небрежны в своих рассуждениях, что же касается необходимой для гёделевского контрдоказательства последовательности выражения мысли, то такое, напротив, встречается далеко не так часто, как хотелось бы.

Следует, впрочем, понимать, что ошибки, подобные тем, что допускали в вышеупомянутых экспериментах студенты, не имеют ничего общего с главным предметом настоящего исследования. Такие ошибки принадлежат к категории «исправимых ошибок» — сами же студенты, несомненно, признáют, что они ошиблись, если им на эти ошибки указать (и, при необходимости, доходчиво разъяснить их природу). Исправимые ошибки мы в данном контексте не рассматриваем вовсе; см., в частности, комментарий к возражению **Q13**, а также §§ 3.12, 3.17. Исследование ошибок, которым порой подвержены люди, безусловно имеет огромное значение для психологии, психиатрии и физиологии, однако меня здесь интересуют совсем другое — а именно, то, что человек может воспринять *в принципе*, используя свои понимание, интуицию и способность к умозаключениям. Как выяснилось, связанные с этим вопросы весьма тонки, хотя тонкость их сразу в глаза не бросается. Поначалу такие вопросы выглядят тривиальными; действительно, корректное рассуждение есть корректное рассуждение, с какой стороны его ни разглядывай, — всего лишь нечто более или менее очевидное, причем все методы такого рассуждения разложил по полочкам еще Аристотель 2300 лет назад (ну а если не он, то английский математик и логик Джордж Буль в 1854 году вкупе с многочисленными последователями). И все же приходится признать, что понятие «корректного рассуждения» таит в себе неизмеримые глубины и совершенно

не укладывается в рамки вычислительных операций, что, в сущности, и показали Гёдель с Тьюрингом. В недавнем прошлом эти вопросы рассматривались как прерогатива скорее математики, чем психологии, присущие же им тонкости психологов в общем случае не интересовали. Однако, как мы могли убедиться, только так можно получить хоть какую-то информацию о физических процессах, которые в конечном счете и обусловливают осознание и понимание.

Исследование упомянутых материй, помимо прочего, неизбежно затронет и глубинные вопросы философии математики. Происходит ли при математическом понимании своего рода контакт с платоновой математической реальностью, существующей независимо от человека и вне времени; или каждый из нас в процессе прохождения этапов логического умозаключения самостоятельно воссоздает все математические концепции? Почему физические законы, как нам представляется, столь неукоснительно следуют полученным таким образом точным и тонким математическим описаниям? Какое отношение имеет собственно физическая реальность к упомянутой концепции платоновой идеальной математической реальности? И, кроме того, если наше восприятие в силу своей природы действительно обусловлено некоей точной и тонкой математической подструктурой, на которую опираются те самые законы, что регулируют функциональную деятельность нашего мозга, то что мы можем узнать о том, как работает наше восприятие математики — как *вообще* работает наше восприятие чего бы то ни было, — если нам удастся глубже понять упомянутые физические законы?

В конечном счете, все наши усилия сводятся к поискам ответов именно на эти вопросы, и к этим же вопросам нам еще предстоит вернуться в конце второй части.

## Примечания

1. Цитата приводится по [329] и [376]. Она, судя по всему, является частью Гиббсовских лекций Гёделя, прочитанных в 1951 году; полный текст имеется в Собрании сочинений Гёделя, том 3 [160]. См. также [377], с. 118.
2. См. [198], с. 361. Цитата взята из лекции Тьюринга, прочитанной в 1947 году перед Лондонским математическим обществом и приводится по изданию [370].



3. Упомянутая процедура заключается во вложении системы $\mathbb{ZF}$ в систему Гёделя–Бернайса; см. [56], глава 2.
4. См. [181], с. 74.
5. Это самое количество состояний Вселенной (число порядка $10^{10^{123}}$ или около того) представляет собой объем доступного фазового пространства (измеренный в абсолютных единицах из § 6.11) некоторой области, содержащей в себе такое количество вещества, какое заключено внутри наблюдаемой нами в настоящий момент Вселенной. Величину этого объема можно оценить, применив формулу Бекенштейна–Хокинга для энтропии черной дыры с массой, равной массе упомянутого количества вещества, и найдя экспоненту от этой энтропии (в абсолютных единицах из § 6.11). См. НРК, с. 340–344.
6. См. [267], [268].
7. См., напр., [102] (и НРК, глава 9).
8. Популярно об этих исследованиях рассказано в [153] и [337].
9. Из классической теории фон Неймана и Моргенштерна (1944).
10. См. [153], [337].
11. Популярное изложение этих вопросов можно найти в [350], [351] и [329].
12. Гипотеза Тебо — это весьма занимательная (и даже не слишком сложная) теорема из плоской евклидовой геометрии, которую, тем не менее, не так-то просто доказать непосредственно. Как выяснилось, единственный способ ее доказательства заключается в том, чтобы отыскать подходящее обобщение (что сделать не в пример легче), а уже затем выводить требуемый результат в виде особого случая. Такая процедура довольно широко распространена в математике, однако для компьютеров она, как правило, совершенно не годится, поскольку отыскание необходимого обобщения требует немалой изобретательности и способности разбираться в сути проблемы. Компьютерное же доказательство подразумевает наличие некоей четкой системы нисходящих правил, которым машина в дальнейшем и следует неуклонно с поражающей воображение скоростью. В данном случае львиная доля человеческой изобретательности как раз и пошла в первую очередь на разработку эффективной системы таких нисходящих правил.
13. Исторический обзор некоторых таких попыток можно найти у Д. Фридмана [124].
14. Это заявление следует рассматривать с учетом сказанного в § 1.8; оно опирается на общепринятое допущение, согласно которому

аналоговые системы можно без особого ущерба для точности рассматривать с помощью численных методов. См. также источники, указанные в примечании 12 после главы 1.

15. Предположение о том, что нейроны представляют собой нечто большее, чем просто «двухпозиционные переключатели», как считалось раньше, похоже, находит поддержку в самых широких научных кругах. См., например, книги Скотта [339], Хамероффа [183], Эдельмана [111] и Прибрама [319]. Как мы увидим в главе 7, некоторые идеи Хамероффа оказываются в нашем контексте чрезвычайно значимыми.
16. См. статьи Г. Фрелиха [129], [130], [131], [132], [133]; дальнейшее развитие эти идеи получили в работах Маршалла [258], Локвуда [243], Зохара [397] и др. В нашем исследовании они также сыграют немаловажную роль; см. § 7.5 и [18].
17. См., например, [346], [316], [29] и [328].
18. Замечательные описания игры Конуэя «Жизнь» можно найти в [137], [311] и [391].
19. См., например, [214] и [40].
20. Подробное описание этих экспериментов приведено в [40].

# Часть II

# НОВАЯ ФИЗИКА, НЕОБХОДИМАЯ ДЛЯ ПОНИМАНИЯ РАЗУМА

## В поисках невычислительной физики разума

# 4

# ЕСТЬ ЛИ В КЛАССИЧЕСКОЙ ФИЗИКЕ МЕСТО РАЗУМУ?

## 4.1. Разум и физические законы

Все мы (как телом, так и разумом) принадлежим Вселенной, которая беспрекословно подчиняется — причем с чрезвычайно высокой точностью — невероятно хитроумным и повсеместно применимым математическим законам. В рамках современного научного мировоззрения уже давно принимается как данность тот факт, что физическое тело человека находится с упомянутыми законами в полном согласии. А разум? Многим глубоко неприятна мысль о том, что нашим разумом управляют все те же математические законы. И все же если нам придется проводить четкую границу между телом и разумом — первое подвержено действию математических законов физики, а второму дозволено быть от них свободным, — то неприятность никуда не денется, а лишь сменит название. Разум человека, вне всякого сомнения, оказывает влияние на то, как именно действует его тело, а физическое состояние этого самого тела не может, в свою очередь, не влиять тем или иным образом на разум. Сама концепция разума, не предполагающая способности разума хоть как-то воздействовать на собственное тело или испытывать какое-либо воздействие с его стороны, представляется довольно бессмысленной. Более того, если разум — не более чем «эпифеномен» (то есть некое явление, неразрывно связанное с физическим состоянием мозга, но совершенно пассивное), побочный продукт деятельности тела, никак на это тело не влияющий, то получается, что разуму отводится роль беспомощного и бесполезного созерцателя. Если же разум способен повлиять на свое материальное тело таким образом,

что тело сможет действовать вопреки законам физики, то под угрозой оказывается точность и общая применимость этих законов. Таким образом, придерживаться в данном случае целиком и полностью «дуалистической» точки зрения (согласно которой законы, управляющие разумом и телом, никак между собой не связаны и друг от друга не зависят) весьма и весьма непросто. Даже если предположить, что управляющие действиями тела физические законы допускают некоторую свободу, в рамках которой разум может каким-то образом влиять на поведение тела, то тогда и сама эта свобода в данном конкретном проявлении должна являться немаловажной составной частью вышеупомянутых физических законов. Неважно, какие именно законы управляют деятельностью разума и с помощью каких средств мы будем эту деятельность описывать, — все они непременно должны являться неотъемлемой частью того грандиозного механизма, что управляет всеми прочими *материальными* проявлениями нашей Вселенной.

На это нам скажут(1), что если мы будем рассматривать «разум» просто как очередную вещественную сущность — пусть даже отличную от обычной материи и построенную на иных принципах, — то совершим, ни много ни мало, «категориальную ошибку». А в качестве доказательства приведут аналогию, в соответствии с которой материальное тело сравнивается с физическим компьютером, а разум — с компьютерной программой. В самом деле, подобные аналогии порой оказываются весьма конструктивными — там, где они уместны, и, безусловно, в тех случаях, когда очевиден риск возникновения путаницы между концепциями разного уровня, необходимо что-то предпринимать. Тем не менее, одного лишь указания на возможную «категориальную ошибку» явно недостаточно для того, чтобы разрешить вполне реальную проблему взаимоотношений разума и тела.

Кроме того, между некоторыми физическими концепциями и в самом деле можно установить равенство, хотя на первый взгляд может показаться, что при этом неизбежно возникает нечто вроде категориальной ошибки. Примером может послужить знаменитая формула Эйнштейна $E = mc^2$, которая устанавливает эффективное равенство энергии и массы. Налицо явная категориальная ошибка — масса есть мера вещественных, материальных объектов, тогда как энергией, как правило, называют несколько туманную абстрактную величину, которая характеризует потенциальную способность к выполнению работы. И все же формула Эйнштейна, связывающая эти две концепции, по сей день остается краеугольным камнем современной физики, а ее справедливость была неоднократно подтверждена экспериментально на примере самых разных физических процессов. Еще более поразительный пример мнимой категориальной ошибки в физике возникает в связи с концепцией *энтропии* (см. например, НРК, глава 7). Определение энтропии крайне субъективно, поскольку она представляет собой, в сущности, лишь некий придаток к понятию «информация»; в то же время энтропия оказывается связана и с другими, более «материальными» физическими величинами посредством вполне точных математических соотношений(2).



Равным образом, я не вижу причин, способных запретить нам хотя бы попытаться рассмотреть концепцию «разума» с точки зрения возможности ее наглядного соотнесения с другими физическими концепциями. В частности, понятие разума непременно должно включать в себя «сознание», неразрывно связанное с вполне определенными и весьма специфическими физическими объектами (с живым и бодрствующим человеческим мозгом, по меньшей мере), так что можно предположить, что какое-никакое физическое описание этого феномена окажется в конечном счете возможным; при этом совершенно неважно, насколько далеки мы от его понимания в настоящий момент. Один шаг к такому пониманию мы сделали в первой части книги: сознательное понимание должно, помимо прочего, сопровождаться некоей неалгоритмической физической активностью, — если, конечно, следовать логике представленных рассуждений и умозаключений, т. е. если мы готовы принять точку зрения, сходную, скорее, с $\mathscr{C}$ (ради чего, собственно, я все это и затеял), нежели с любой из остальных ($\mathscr{A}$, $\mathscr{B}$ и $\mathscr{D}$, см. § 1.3). Я прошу тех читателей, кого не убедили мои предыдущие аргументы, не покидать нас еще некоторое время и хотя бы взглянуть на те неведомые края, к исследованию которых нас побуждает $\mathscr{C}$. Мы обнаружим, что открывающиеся перед нами возможные варианты вовсе не так бесперспективны, как, казалось бы, можно было ожидать; многое в этих краях и само по себе представляет немалый интерес. Надеюсь, что по завершении наших изысканий упомянутые читатели с большей благосклонностью отнесутся к предложенным в первой части книги аргументам (и оценят, наконец, их красоту и мощь). Отправимся же в путь — вслед за нашей путеводной звездой $\mathscr{C}$!

## 4.2. Вычислимость и хаос в современной физике

Точность и область применимости физических законов, по современным оценкам, чрезвычайно велики, однако в этих законах нет ни единого намека на процессы, которые невозможно моделировать вычислительными методами. Тем не менее, мы все же попробуем отыскать в дозволенных законами пределах место для той таинственной невычислительной активности, которая каким-то образом оказывается необходимой для функционирования наших с вами мозгов. Отложим на некоторое время дискуссию о возможной природе такой невычислимости. Есть все основания полагать, что природа эта чрезвычайно хитроумна и неуловима, и мне бы не хотелось застрять в самом начале, увязнув в рассмотрении всех непременно связанных с нею тонкостей. Мы вернемся к этому вопросу позже (§§ 7.9, 7.10). Достаточно сказать, что для хоть какого-то движения вперед нам потребуется нечто существенно отличное от тех картин, что рисуют существующие на данный момент физические теории, будь они классическими или квантовыми.

В *классической* физике мы можем в любой выбранный момент времени указать все необходимые для определения физической системы данные, дальнейшая же эволюция этой системы не только целиком и полностью определяется указанными данными, но и может быть по ним *вычислена* с помощью эффективных методов «тьюрингова» вычисления. По крайней мере, такое вычисление возможно *в принципе*, при соблюдении двух взаимосвязанных условий. Первое условие заключается в возможности адекватной оцифровки исходных данных — с тем, чтобы мы могли с достаточной степенью точности заменить непрерывные параметры теории соответствующими *дискретными* параметрами. (В сущности, такая замена обычно и производится при компьютерном моделировании классических систем.) Второе условие связано с тем фактом, что многие физические системы являются *хаотическими* — в том смысле, что вычисление дальнейшего поведения такой системы с хоть сколько-нибудь приемлемой точностью требует совершенно непомерной точности исходных данных. Выше (см., в частности, § 1.7, а также §§ 3.10, 3.22) мы уже рассмотрели такие системы довольно подробно и пришли к выводу, что хаотическое поведение в дискретно действующей системе *не* приводит к той «невычислимости», которая нас в данном случае интересует. Хаотическая (дискретная) система, пусть



и сложная для вычисления, остается все же системой вычислимой, о чем свидетельствует тот факт, что подобные системы, как правило, исследуются и моделируются посредством электронных компьютеров! Первое условие связано со вторым, поскольку в хаотической системе ответ на вопрос о том, какую степень точности дискретной аппроксимации к непрерывным параметрам теории следует полагать «адекватной», зависит от того, намерены мы вычислять *действительное* поведение системы или достаточно будет и *типичного*. Если только последнее (а как я показал в первой части, большего, коль скоро речь идет об искусственном интеллекте, по всей видимости, и не требуется), то нет нужды беспокоиться о том, что наши дискретные аппроксимации окажутся несовершенными, а малые погрешности в исходных данных приведут к огромным отклонениям в последующем поведении системы. Если нас и в самом деле занимает лишь типичное поведение, то вышеприведенные условия не оставляют места для сколько-нибудь серьезной возможности возникновения в любой чисто классической физической системе невычислимости требуемого (в соответствии с рассуждениями, представленными в первой части книги) рода.

Не следует, впрочем, сбрасывать со счетов возможности наличия в действительном хаотическом поведении какой-нибудь непрерывной математической системы (моделирующей некое реальное физическое поведение) процессов, воспроизвести которые с помощью дискретной аппроксимации *в принципе* невозможно. Я ни о чем подобном никогда не слышал, однако даже если такая система где-нибудь и существует, создателям искусственного интеллекта (в том виде, как мы понимаем его сегодня) от нее никакого проку не будет, поскольку все современные разработки в этой области опираются как раз на *дискретное* вычисление (т. е. на вычисление скорее цифровое, нежели аналоговое; см. § 1.8).

В *квантовой* физике, наряду с детерминированным (и вычислимым) поведением, описываемым уравнениями квантовой теории (в основном, уравнением Шрёдингера), присутствует и некая добавочная степень свободы, целиком и полностью *случайная* по своей природе. С формальной точки зрения, уравнения квантовой теории *не* являются хаотическими, однако отсутствие хаоса возмещается наличием вышеупомянутых случайных ингредиентов, дополняющих детерминистскую эволюцию. Как мы могли убедиться (в частности, в § 3.18), такие чисто случай-

ные ингредиенты также не в состоянии обусловить необходимую неалгоритмическую активность. Таким образом, ни в классической, ни в квантовой физике (в их теперешнем понимании) для невычислительного поведения требуемого типа просто нет места, поэтому если нам нужна именно невычислительная активность, то искать ее следует где угодно, но только не здесь.

## 4.3. Сознание: новая физика или «эмергентный феномен»?

В первой части я показал (на конкретном примере математического понимания), что феномен *сознания* возникает лишь при условии протекания в мозге неких физических процессов невычислительного характера. Следует, впрочем, допустить, что подобные гипотетические невычислительные процессы должны протекать *и* в неодушевленной материи, поскольку живой человеческий мозг, в конечном счете, из этой самой материи и состоит и подчиняется тем же физическим законам, каким подчиняются все неодушевленные объекты во Вселенной. Таким образом, перед нами встают два вопроса. Первый: почему феномен сознания проявляется, насколько нам известно, *лишь* в мозге (или в той или иной связи с мозгом) — при том, что полностью исключить возможность присутствия сознания и в других достаточно сложных физических системах нельзя? И второй вопрос: чем объяснить тот факт, что такой, казалось бы, важный (пусть и гипотетический) ингредиент, как невычислительное поведение, — к тому же непременно, согласно нашему допущению, присутствующий (по крайней мере, потенциально) в физической активности всех материальных объектов — умудрился ни разу до сих пор не попасться на глаза физикам?

Ответ на первый вопрос, несомненно, имеет какое-то отношение к сложной и изощренной организации мозга, однако какой бы ни была эта организация, сама по себе она еще не может служить достаточным объяснением. Согласно выдвигаемым мною здесь идеям, организация мозга происходит из необходимости реализации невычислительной активности в рамках физических законов; прочая же материя в подобной организации не нуждается. Эта картина разительно отличается от более общепринятого (совпадающего, по большей части, с точкой зрения $\mathscr{A}$) взгляда на природу сознания[3], в соответствии с которым осмысленное осознание представляет собой своего рода «эмергентный феномен», т. е. свойство системы, естественным образом возникающее по достижении этой системой достаточной степени организационной и функциональной сложности и не требующее для своего возникновения запуска каких-то новых фундаментальных физических процессов, принципиально отличных от тех, что уже известны из наблюдений за поведением неодушевленной материи. В первой части я пришел к иному выводу: для возникновения сознания одной лишь сложности мало, мозг должен быть организован именно так, чтобы в нем могли протекать предполагаемые невычислительные физические процессы. Более детальные комментарии относительно возможной природы такой организации я приведу позже (§§ 7.4–7.7).



Что касается второго вопроса, то, действительно, следует предположить, что следы интересующей нас невычислимости непременно должны присутствовать (на некоем неразличимом уровне) и в неодушевленной материи. Однако физика «обыкновенной» материи не оставляет (по крайней мере, на первый взгляд) места для такого невычислительного поведения. В дальнейшем я попытаюсь объяснить подробнее, каким образом это невычислительное поведение могло остаться незамеченным и как оно согласуется с современными наблюдениями. Пока же, думаю, будет полезно рассмотреть один феномен из уже *известной* физики — совершенно посторонний, но не лишенный некоторых весьма близких аналогий. Хотя данный физический феномен не связан (*непосредственно*, по крайней мере) с каким бы то ни было невычислительным поведением, он очень похож на наш гипотетический невычислимый ингредиент в ином отношении — его совершенно невозможно обнаружить даже при тщательном наблюдении поведения обыкновенных объектов. На соответствующем уровне он, впрочем, проявляется и, как выяснилось, коренным образом изменяет наше представление о том, как устроен мир, — по сути определяя тем самым дальнейшее направление развития науки в целом.

## 4.4. Эйнштейнов *наклон*

Со времен Исаака Ньютона и до наших дней физический феномен *гравитации* — вместе с замечательно точным математическим его описанием (впервые представленным Ньютоном

в полном виде в 1687 году) — играет в развитии научной мысли одну из ключевых ролей. После окончательного утверждения математического аппарата гравитация могла служить (и послужила) прекрасной моделью для описания самых разных физических процессов; при этом предполагалось, что движения тел в неподвижном (плоском) опорном пространстве точно определяются действующими на эти тела силами — силами взаимного притяжения (либо отталкивания) отдельных частиц, управляющими любым движением этих частиц, вплоть до самого незначительного. Результатом выдающегося успеха ньютоновской теории тяготения стала постепенно укрепившаяся вера в то, что таким образом можно описать вообще *все* физические процессы, — исходя из предположения, что электрические, магнитные, молекулярные и прочие силы точно так же действуют между частицами и так же, в общем, управляют их мельчайшими движениями, как и силы гравитационные.

Некое возмущение в эту идиллическую картину внес в 1865 году великий шотландский физик Джеймс Клерк Максвелл, опубликовав свою знаменитую систему уравнений, точно описывающую поведение электрических и магнитных полей. Теперь, наряду с всевозможными дискретными частицами, пришлось признать независимое существование и этих непрерывных полей. Электромагнитное поле (как называют сегодня комбинацию двух упомянутых полей) способно осуществлять перенос энергии через в прочем отношении пустое пространство — в виде света, радиоволн, рентгеновских лучей и т. д. — и ничуть не менее реально, чем ньютоновские частицы, с которыми оно, как предполагается, сосуществует. Тем не менее, объектом общего описания и здесь остаются физические тела (к каковым теперь причисляются и непрерывные поля), движущиеся в неподвижном пространстве в результате неких взаимодействий друг с другом, т. е. в общем и целом ньютоновская схема существенных изменений не претерпела. Даже вводимая в 1913–1926 годах стараниями Нильса Бора, Вернера Гейзенберга, Эрвина Шрёдингера, Поля Дирака и др. квантовая теория, со всей ее революционностью и эксцентричностью, не изменила этого аспекта нашего физического мировоззрения. Физические объекты продолжали восприниматься как некие сущности, действующие друг на друга посредством силовых полей, причем и те, и другие пребывали все в том же неподвижном, плоском, опорном пространстве.



В годы появления первых работ в области квантовой теории Альберт Эйнштейн был занят тем, что подвергал глубокому пересмотру сами фундаментальные основы ньютоновской теории тяготения, результатом чего стала представленная им в 1915 году революционно *новая* теория, совершенно изменившая привычную картину мира, — речь идет, конечно же, об общей теории относительности (см. НРК, с. 202–211). Гравитация здесь вообще не является силой, ее следует представлять как своего рода *искривление* самого пространства (в действительности, даже пространства–времени), в которое помещаются все прочие частицы и силы.

Далеко не всем физикам эта «несообразная» картина пришлась по душе. Им не понравилось, что гравитация оказалась в таком отрыве от остальных физических воздействий, — особенно принимая во внимание тот факт, что именно гравитация послужила основой для первоначальной парадигмы, по образу и подобию которой были выстроены все более поздние физические теории. Еще одним поводом для недоверия стало то, что гравитационное взаимодействие чрезвычайно слабó — в сравнении с прочими известными физикам силами. Например, сила гравитационного притяжения между электроном и протоном в атоме водорода в

28 500 000 000 000 000 000 000 000 000 000 000 000 000

раз меньше, чем сила электрического взаимодействия между этими же частицами. То есть на уровне отдельных частиц, составляющих материю, гравитационные силы практически незаметны.

Не раз поднимался вопрос о том, не является ли гравитация своего рода *остаточным* эффектом, этаким *последействием*, возникающим, скажем, при почти полной взаимной компенсации всех сил, действующих в данной системе? (Такие силы в природе действительно существуют — например, сила Ван-дер-Ваальса, водородная связь и сила Лондона.) При таком подходе перед нами оказывается уже не самостоятельный физический феномен, отличный от всех прочих и нуждающийся поэтому в совершенно особом (отличном от описания всех прочих сил) математическом описании, — при таком подходе гравитации как таковой в действительности не существует, а существует лишь своего рода «эмергентный феномен». (Подобный взгляд на гравитацию предложил великий советский ученый и гуманист Андрей Сахаров[4].)

Впрочем, как выяснилось позднее, такое предположение *лишено* оснований. Главная причина заключается в том, что гравитация воздействует на причинные связи между пространственно-временны́ми событиями; *никакая* другая физическая величина такого воздействия не производит. Можно сказать иначе: гравитация обладает уникальной способностью «наклонять» световые конусы. (Вскоре я поясню, что все это означает.) *Только* гравитация может наклонять световые конусы, *никакая* другая физическая сила (равно как и *никакая* комбинация любых негравитационных физических воздействий) на это не способна.

Что же означает выражение «наклон светового конуса»? Что такое «причинные связи между пространственно-временны́ми событиями»? Для объяснения этих терминов нам потребуется несколько отклониться от темы. (Это отклонение еще сослужит нам в дальнейшем хорошую службу.) Некоторые читатели, возможно, уже знакомы с соответствующими научными концепциями, поэтому я дам здесь лишь краткое описание — с тем, чтобы и остальные могли получить хоть какое-то представление о предмете. (См. также НРК, глава 5, с. 194, там все рассмотрено более подробно.) На рис. 4.1 я изобразил единичный световой конус в пространственно-временны́х координатах. Ось времени на рисунке направлена снизу вверх, пространство же «откладывается» по горизонтали. Точкой на пространственно-временно́й диаграмме отображается *событие*, т. е. некая точка пространства в какой-то определенный момент времени. Событие, таким образом, имеет нулевую временну́ю продолжительность, равно как и нулевую пространственную протяженность. Полный *световой конус* с центром в точке-событии $P$ представляет пространственно-временну́ю историю сферического светового импульса, который «схлопывается» внутрь $P$ и тут же «выплескивается» обратно, наружу; все это, разумеется, происходит со скоростью света. Таким образом, световой конус события $P$ образуют все те лучи света, в индивидуальной истории которых событие $P$ происходило.

Световой конус $P$ состоит из двух частей: светового конуса *прошлого*[1] (*входящая* вспышка) и светового конуса *будущего* (*исходящая* вспышка). Согласно теории относительности, причинное воздействие на пространственно-временное событие $P$

[1] На рисунках в НРК изображены только «будущие» части световых конусов.



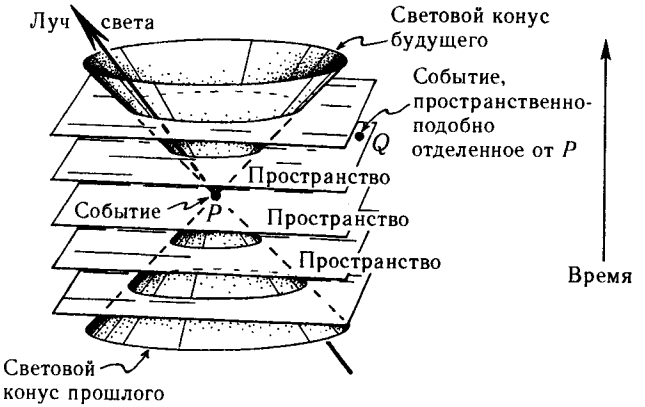



Рис. 4.1. *Световой конус* события $P$ составляют все те лучи света, которые в пространстве-времени проходят через событие $P$. Сам конус представляет собой историю вспышки света, схлопывающейся в точку $P$ (световой конус прошлого) и вырывающейся затем наружу (световой конус будущего). События $Q$ и $P$ *пространственноподобно разделены* (точка $Q$ лежит вне светового конуса $P$), т. е. событие $Q$ оказывается вне зоны причинного воздействия события $P$.

способны оказать только события, расположенные либо внутри светового конуса прошлого $P$, либо на его поверхности; аналогично, само событие $P$ способно оказать причинное воздействие только на те события которые расположены либо внутри светового конуса будущего $P$, либо на его поверхности. События, расположенные вне световых конусов прошлого и будущего, не могут ни воздействовать на событие $P$, ни подвергаться воздействию со стороны события $P$. Мы говорим, что такие события *пространственноподобно* отделены от $P$.

Следует помнить, что понятие причинной связи принадлежит теории *относительности*; к ньютоновской физике оно никакого отношения не имеет. В ньютоновской картине мира скорость передачи информации ничем не ограничена. В теории же относительности у этой скорости появляется предел — скорость света. Отсюда один из фундаментальных принципов теории от-

носительности: никакое причинно-следственное воздействие не может происходить со скоростью, превышающей скорость света.

Впрочем, при толковании термина «скорость света» нужно соблюдать известную осторожность. Реальные световые сигналы несколько замедляются при прохождении через преломляющую среду (такую, например, как стекло). В такой среде скорость распространения физического светового сигнала будет меньше, чем скорость, которую мы здесь называем «скоростью света», и вполне возможно, что какое-либо физическое тело (или сигнал, отличный от светового) будет здесь перемещаться быстрее света. Этот феномен можно наблюдать в некоторых физических экспериментах (например, экспериментах по получению так называемого черенковского излучения). Частицы «выстреливаются» в преломляющую среду, в которой скорость частиц лишь очень немногим меньше абсолютной «скорости света», но больше скорости, с которой свет фактически распространяется в данной среде. При этом возникают ударные волны «реального» света, которые и называются черенковским излучением.

Во избежание путаницы я лучше буду называть бóльшую «скорость света» *абсолютной* скоростью. Световые конусы в пространстве-времени определяют абсолютную скорость, но эта скорость совсем не обязательно равна действительной скорости света в каждом конкретном случае. Внутри какой-либо среды действительная скорость света несколько меньше абсолютной скорости, равно как и меньше скорости перемещающихся в этой среде частиц, генерирующих черенковское излучение. Пределом же скорости как для сигналов, так и для материальных тел является именно абсолютная скорость (оба световых конуса), и хотя реальный свет отнюдь не всегда распространяется с абсолютной скоростью, в вакууме скорость света совпадает с абсолютной.

Теорию «относительности», о которой мы здесь в основном говорим, называют еще *специальной* теорией относительности — специальной, поскольку в ней не учитывается гравитация. Все световые конусы в специальной теории относительности размещены равномерно и сориентированы в одном направлении (как показано на рис. 4.2); такое пространство-время называют *пространством Минковского*. Согласно же *общей* теории относительности Эйнштейна, предыдущие рассуждения остаются в силе только если мы продолжаем считать «абсолютной» ту скорость, что определяется пространственно-временным положением све-



товых конусов. Однако под воздействием гравитации распределение световых конусов может стать *неоднородным* (рис. 4.3). Именно это я и подразумевал, говоря выше о «наклоне» световых конусов.

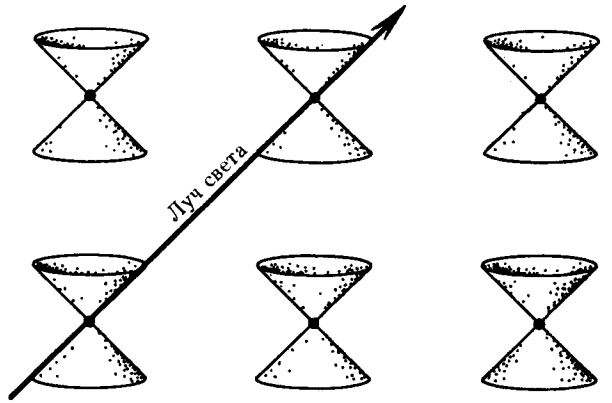



Рис. 4.2. Пространство Минковского: пространство-время в специальной теории относительности. Все световые конусы размещены равномерно и сориентированы в одном направлении.

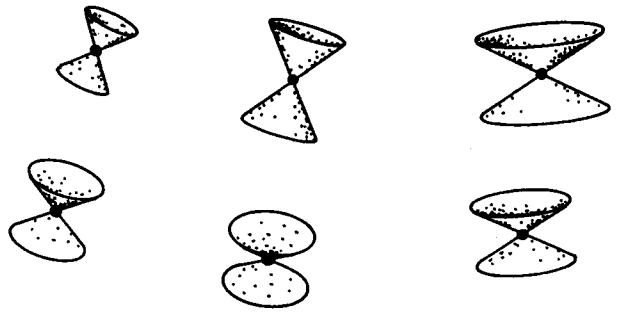


Рис. 4.3. *Наклонные* световые конусы в общей теории относительности Эйнштейна.

Наклон световых конусов можно представлять себе как *изменение* скорости света (или, точнее, абсолютной скорости) в зависимости от места в пространстве; эта скорость может также зависеть и от направления движения. При таком подходе «абсолютную скорость» можно рассматривать как некий аналог «действительной скорости света» в преломляющих средах, о которой мы говорили выше. Соответственно, можно предположить, что гравитационное поле является этакой всепроницающей и повсеместной преломляющей средой, которая оказывает воздействие не только на поведение реального света, но и на поведение *всех* материальных частиц и сигналов[2]. В самом деле, попытки описать феномен и эффекты гравитации именно таким образом предпринимаются нередко, и до некоторой степени это описание работает. Однако в общем и целом это описание оказывается неудовлетворительным, а в некоторых существенных отношениях и вовсе дает серьезно искаженную картину общей относительности.

Прежде всего следует отметить, что хотя такую «гравитационную преломляющую среду» и можно счесть причиной *уменьшения* абсолютной скорости (как обстоит дело с обычной преломляющей средой), некоторые существенные обстоятельства (например, большая протяженность гравитационного поля изолированной массы) не позволяют ограничиться одним лишь *замедляющим* воздействием — кое-где наша гипотетическая среда должна проявить способности и к воздействию *ускоряющему*, т. е. где-то абсолютная скорость должна *возрастать* (см. [290] и рис. 4.4). В рамках специальной теории относительности *такое* просто *невозможно*. Согласно этой теории, никакая преломляющая среда, сколь бы причудливой она ни была, не может разгонять сигналы до скорости, превышающей скорость света в вакууме (т. е. в отсутствие какой бы то ни было среды), не нарушая при этом фундаментальных для теории принципов причинности — ведь такое увеличение скорости позволило бы сигналам распространяться снаружи минковскианских световых конусов (вакуумных), а это теоретически запрещено. К тому же, как мы выяснили выше, гравитационные эффекты «наклона световых конусов» нельзя объяснить никаким остаточным воздействием прочих, негравитационных, полей.

[2]Забавно, что сам Ньютон тоже высказывал подобную идею. (См. «Вопросы» 18–22 в третьей книге «Оптики» (1730).)



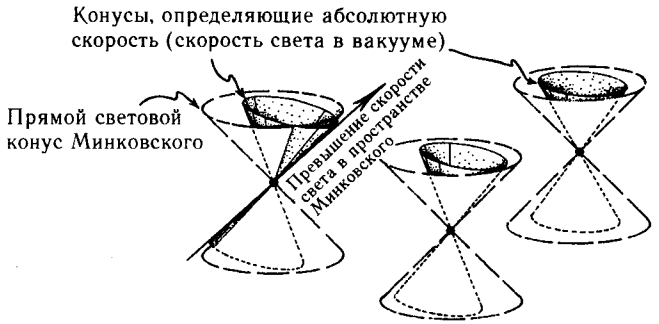



Рис. 4.4. Распространение света согласно общей теории относительности Эйнштейна не может являться эффектом «преломляющей среды» (в пространстве Минковского), поскольку это противоречит фундаментальному принципу специальной теории относительности — невозможности распространения сигналов со скоростью, превышающей скорость света в пространстве Минковского.

Известны и гораздо более «экстремальные» ситуации, в которых описать таким образом наклон световых конусов и вовсе невозможно, даже если допустить «превышение» абсолютное скорости в некоторых направлениях. Одну такую ситуацию иллюстрирует рис. 4.5: световые конусы наклонены под самым невероятным углом, чуть ли не перевернуты. Вообще говоря, такой чрезвычайный наклон возникает лишь в явно спорных ситуациях, где имеет место так называемое «нарушение причинности» — т. е. наблюдатель получает теоретическую возможность посылать сигналы в свое собственное прошлое (см. рис. 7.15, глава 7). Отметим еще, что соображения такого рода, как это ни удивительно, имеют самое что ни на есть непосредственное отношение к одной из тем нашего дальнейшего обсуждения (см. § 7.10).

Следует упомянуть и еще об одном неявном обстоятельстве: «угол наклона» единичного светового конуса не является величиной, измеримой физически, а потому не имеет в сущности никакого физического смысла и не может послужить мерой *действительного* уменьшения или увеличения абсолютной скорости. Лучшим способом проиллюстрировать это обстоятельство

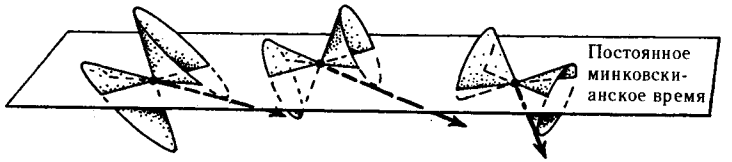



Рис. 4.5. В принципе наклон светового конуса может стать настолько большим, что сигналы смогут распространяться в минковскианское прошлое.

будет следующий: вообразим, что изображение, представленное на рис. 4.3, нанесено на тонкий лист резины, что позволит поворачивать и деформировать каждый отдельный световой конус вокруг окрестности его вершины (см. рис. 4.6) до тех пор, пока он не расположится «вертикально», — т. е. так, как располагаются световые конусы в пространстве специальной относительности Минковского (рис. 4.2). При этом нет никакой возможности обнаружить (посредством локальных экспериментов), является ли «наклонным» световой конус того или иного конкретного события. Если же мы намерены настаивать на том, что «эффект наклона» обязан своим возникновением некоей «гравитационной среде», то нам придется объяснить и «странности» поведения этой самой среды — объяснить, почему эта среда ни при каком единичном пространственно-временном событии не поддается наблюдению. В частности, даже очевидно чрезвычайные случаи (представленные на рис. 4.5), для описания которых идея гравитационной среды ну совершенно не годится, оказываются неотличимы физически (если рассматривать один-единственный световой конус) от случая, когда наклон отсутствует (как в пространстве Минковского).

Впрочем, если говорить вообще, то поворачивать тот или иной конкретный световой конус до его минковскианской ориентации мы можем лишь за счет деформации — и *удаления* от минковскианской ориентации — некоторых из соседних световых конусов. Возникает, в общем случае, «математическое препятствие», в силу которого невозможно деформировать лист резины таким образом, чтобы все световые конусы выстроились в стандартный минковскианский порядок, показанный на рис. 4.2. В четырехмерном пространстве-времени это препятствие описывается посредством математического объекта, называемого *конформным тензором Вейля* — в НРК мы ввели для этого тензора обозначение **WEYL** (см. НРК, с. 210). (Тензор **WEYL** дает ровно половину — «конформную» половину — информации, содержащейся в полном тензоре пространственно-временно́й кривизны Римана; впрочем, полагаю, что в данной ситуации беспокоиться о точном смысле этих терминов особой необходимости нет.) Развернуть *все* световые конусы в минковскианский порядок нам удастся лишь в том случае, если **WEYL** будет равен нулю. Тензор **WEYL** есть мера гравитационного поля — в смысле гравитационной приливной деформации, — т. е. именно *гравитационное поле* и является тем самым препятствием, которое не дает нам «выпрямить» все световые конусы сразу.



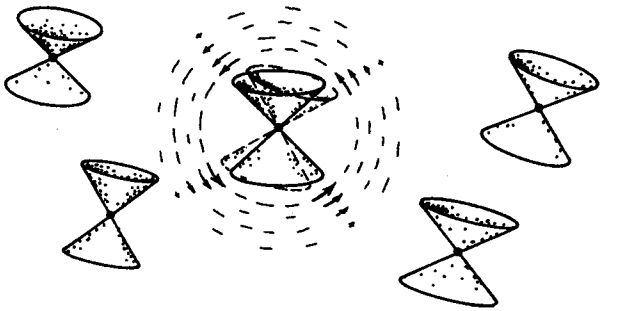


Рис. 4.6. Вообразим пространство-время в виде резинового листа с нанесенными на нем световыми конусами. Каждый отдельный световой конус можно поворачивать (растягивая резину) до тех пор, пока все они не выстроятся в стандартную минковскианскую картину.

Эту тензорную величину, конечно же, можно измерить физически. **WEYL**-тензорное гравитационное поле, например, Луны воздействует на Землю и вызывает ее приливную деформацию — внося тем самым основной вклад в возникновение приливов (см. НРК, с. 204, рис. 5.25). Этот эффект, впрочем, не связан непосредственно с наклоном световых конусов, а представляет собой лишь самое обычное проявление ньютоновского гравитационного воздействия. Более подходящим к случаю выглядит другой наблюдаемый эффект, так называемый *эффект гравитационной*

*линзы*, предсказанный в теории Эйнштейна. Впервые гравитационную линзу наблюдал Артур Эддингтон во время экспедиции на остров Принсипи в 1919 году; при этом вызванное гравитационным полем Солнца искажение картины звездного неба было самым тщательным образом зарегистрировано. Звездное небо вблизи Солнца словно растягивается — при этом, скажем, небольшой круг из звезд представляется наблюдателю в виде эллипса (см. рис. 4.7). В данном случае воздействие **WEYL**-тензорного гравитационного поля на структуру световых конусов пространства-времени наблюдалось почти непосредственно. В последние годы эффект гравитационной линзы находит широкое применение в качестве инструмента наблюдательной астрономии и космологии. Свет от отдаленного квазара порой доходит до нас в искаженном виде, поскольку на его пути оказывается какая-либо крупная масса (например, галактика; см. рис. 4.8). Из наблюдаемых при этом искажений «внешности» квазара (вкупе с эффектами временной задержки) можно извлечь весьма ценные сведения о соответствующих расстояниях, массах и т. д. Все это можно полагать достаточно недвусмысленным свидетельством в пользу того, что феномен наклона световых конусов действительно существует, а также того, что **WEYL**-эффекты непосредственно измеримы.

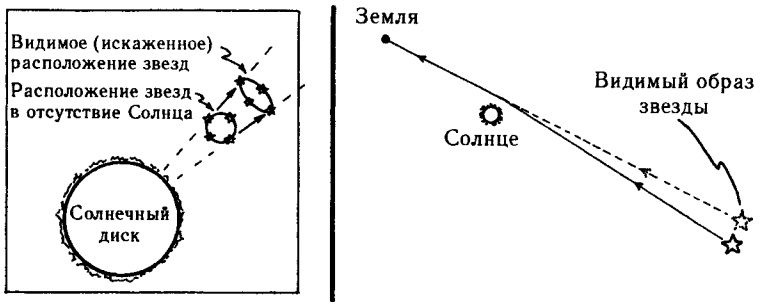



Рис. 4.7. Непосредственно наблюдаемый эффект наклона световых конусов. Пространственно-временно́е **WEYL**-искривление проявляется в виде искажения картины звездного неба в результате отклонения световых лучей под воздействием гравитационного поля Солнца. Круг из звезд представляется наблюдателю эллипсом.



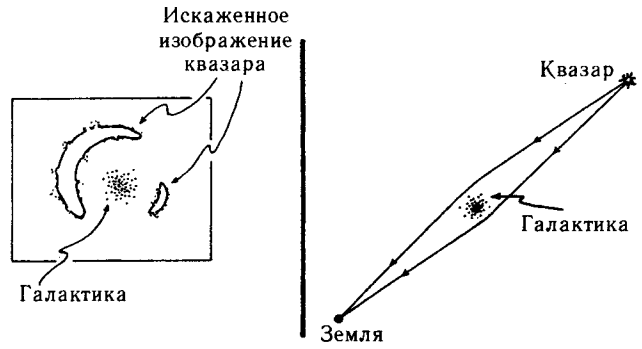



Рис. 4.8. Эффект эйнштейновского отклонения света широко используется сегодня в наблюдательной астрономии. По тому, насколько искажено изображение отдаленного квазара, можно оценить массу галактики, находящейся между квазаром и наблюдателем.

Предыдущие замечания наглядно иллюстрируют тот факт, что «наклон» световых конусов, т. е. гравитационное искажение причинности, представляет собой не нечто эфемерное, но вполне *реальный* феномен, который нельзя исчерпывающе объяснить каким бы то ни было остаточным (либо «эмергентным») свойством, возникающим у достигшего достаточной величины скопления материи. Гравитация имеет собственную *уникальную* природу, отличную от природы прочих физических процессов; на уровне тех сил, что существенны для фундаментальных частиц, гравитация непосредственно не наблюдается — тем не менее, она присутствует и здесь, и присутствует постоянно. Наклон световых конусов — прерогатива гравитации, никакие *другие* из известных современной физике сил и взаимодействий на это не способны. Таким образом, в этом фундаментальном отношении гравитация представляет собой нечто особенное, нечто принципиально *отличное* от всех известных нам сил и физических воздействий. В самом деле, согласно классической общей теории относительности, наклон светового конуса вызывает присутствие любого материального тела, будь оно даже мельчайшей из песчинок (хотя в этом случае наклон будет, конечно же, крайне незначителен). В принципе, для наклона светового конуса достаточно

и отдельного электрона — просто величина производимого подобными объектами наклона слишком мала, чтобы можно было говорить о каком бы то ни было непосредственно наблюдаемом его эффекте.

Гравитационные взаимодействия наблюдались на примере объектов, значительно больших, нежели песчинки, но все же гораздо меньших, чем, например, Луна. В 1798 году Генри Кавендишу удалось измерить силу гравитационного притяжения шара массой всего около $10^5$ граммов. (Этот знаменитый опыт Кавендиша основан на идее, выдвинутой ранее Джоном Мичеллом.) Возможности современной техники позволяют обнаружить гравитационное притяжение объектов значительно менее массивных (см., например, [60]). Впрочем, обнаружить в какой-либо из этих ситуаций эффект наклона световых конусов никакая современная техника пока не в состоянии. Наблюдать этот эффект непосредственно можно только в присутствии действительно огромных масс; а то, что наклон световых конусов создают и малые массы (величиной с песчинку), является очевидным следствием из теории относительности Эйнштейна.

Гравитационные эффекты невозможно сколько-нибудь точно смоделировать посредством какой бы то ни было комбинации других физических полей или сил. Гравитация совершенно уникальна по своей природе, и ни в коем случае нельзя ее рассматривать как эмергентный или вторичный феномен, остаточный по отношению к каким-то иным, более «солидным» физическим процессам. Гравитация описывается самой структурой пространства-времени, которое считалось прежде просто неподвижным фоном, этакой ареной для проявления всевозможной физической активности. В ньютоновской вселенной гравитация не являлась чем-то особенным — хотя и послужила парадигмой для построения всех более поздних физических теорий. Во вселенной же, описываемой Эйнштейном, гравитация рассматривается (и надо сказать, что эта точка зрения, разделяемая большинством нынешних физиков, получила великолепное экспериментальное подтверждение) как совершенно особое взаимодействие — не эмергентный феномен, но нечто само по себе уникальное.

Впрочем, несмотря на все отличия, между гравитацией и прочими физическими силами существует фундаментальная и гармоничная связь. Теория Эйнштейна отнюдь не является чужеродным элементом в системе физических законов, она лишь представляет их в несколько ином свете. (В особенности это относится к законам сохранения энергии, импульса и момента импульса.) Связь эйнштейновской гравитации со всей остальной физикой может до некоторой степени объяснить сложившуюся парадоксальную ситуацию, когда всякое физическое описание основывается на *парадигме* ньютоновской гравитации, в то время как сама гравитация (как позднее показал Эйнштейн) по своей природе *отлична* от прочих физических взаимодействий. Тот же Эйнштейн, кстати, призывал более всего избегать излишней самоуверенности — то, что мы в процессе познания мира взобрались на очередную ступеньку, вовсе не обязательно должно означать, что теперь мы располагаем единственно верной физической теорией этого самого мира.



Можно ли ожидать, что и в отношении феномена сознания нам предстоит обнаружить некое «взаимодействие», аналогичное гравитации? Если да, то характеристикой, которая по достижении определенного значения обусловливает проявление упомянутого феномена, окажется, скорее всего, не *масса* — во всяком случае, не *одна лишь* масса, — но некая разновидность тонкой физической организации. Согласно представленным в первой части доводам, такая организация в процессе своего становления должна была так или иначе научиться использовать некий не известный нам пока ингредиент, непременно присутствующий в поведении обычной материи. То, что мы не наблюдаем его проявлений, означает лишь, что мы не туда смотрим, — аналогичным образом, нам никогда не удалось бы обнаружить феномен наклона световых конусов, ограничь мы область наблюдений одними лишь крохотными частицами.

Какое же отношение имеет наклон световых конусов к невычислимости? К этому вопросу (точнее, к одному весьма интригующему его аспекту) мы еще вернемся в § 7.10; на данном же этапе наших рассуждений ответ прост: абсолютно никакого, *разве что* дает некую надежду — как выясняется, вполне возможно обнаружить в физике фундаментально важное новое свойство, полностью отличное от всех уже известных и остававшееся прежде незамеченным в поведении обычной материи. Эйнштейна к его революционным идеям привел целый ряд весьма мощных соображений — математически сложных и физически неочевидных, — причем самое важное из них, широко известное еще со времен

Галилея, так и оставалось до конца не понятым (речь идет о принципе эквивалентности: все тела в поле тяготения падают с одинаковой скоростью). Более того, необходимое условие успеха идей Эйнштейна заключалось именно в том, что эти самые идеи оказались полностью «совместимыми» со всем тем, что было известно о физических феноменах в его время.

Аналогичным образом вполне можно предположить, что где-то в поведении всем известных объектов сокрыта невычислительная активность того или иного рода. Для того, чтобы подобные спекуляции имели бы хоть какую-то надежду на успех, они также должны быть основаны на каких-то мощных соображениях — предположительно, *и* математически сложных, *и* физически неочевидных — и как-то согласовываться с тем, что мы знаем о всех известных нам феноменах. Посмотрим, насколько далеко нам удастся зайти по пути к такой теории.

Однако прежде чем мы начнем, думаю, стоит составить для себя некоторое представление о том, насколько велико влияние идеи о вычислимости всего и вся на современную физику. Примечательно, что одним из наиболее впечатляющих в этом отношении примеров является не что иное, как общая теория относительности.

## 4.5. Вычисления и физика

На расстоянии около 30 000 световых лет от Земли, в созвездии Орла, есть две невероятно плотные мертвые звезды, вращающиеся одна вокруг другой. Вещество в этих звездах сжато до такой степени, что если сделать из него теннисный мячик, то масса его окажется сопоставима с массой Деймоса, одного из спутников Марса. Время полного оборота этих звезд (называемых обычно ***нейтронными*** звездами) друг вокруг друга составляет 7 часов 45 минут и 6,9816132 секунды, а их массы больше массы Солнца, соответственно, в 1,4411 и 1,3874 раз (с возможной погрешностью в 7 десятитысячных). Каждые 59 миллисекунд первая из этих звезд испускает в нашем направлении импульс электромагнитного излучения (пучок радиоволн), из чего можно заключить, что она вращается вокруг своей оси со скоростью приблизительно 17 оборотов в секунду. Такие звезды называются ***пульсарами***, а описываемая пара звезд представляет собой знаменитый двойной пульсар PSR 1913+16.



Впервые эти замечательные объекты были обнаружены в 1967 году астрономами кембриджской радиообсерватории Джослином Беллом и Энтони Хьюишем. Нейтронные звезды, как правило, являются результатом гравитационного коллапса ядра красного гиганта, каковой коллапс может сопровождаться чрезвычайно яркой вспышкой сверхновой. Нейтронные звезды немыслимо плотны, поскольку состоят из ядерных частиц (в основном, из нейтронов), уложенных настолько близко друг к другу, что общая плотность звезды оказывается сопоставима с плотностью собственно нейтрона. В процессе коллапса нейтронная звезда захватывает своим веществом линии индукции магнитного поля и, вследствие чудовищного сжатия, которым сопровождается коллапс, концентрация этого поля достигает чрезвычайно больших величин. Линии поля выходят из северного магнитного полюса звезды, удаляясь в пространстве на весьма значительное расстояние, и входят в южный магнитный полюс (см. рис. 4.9).

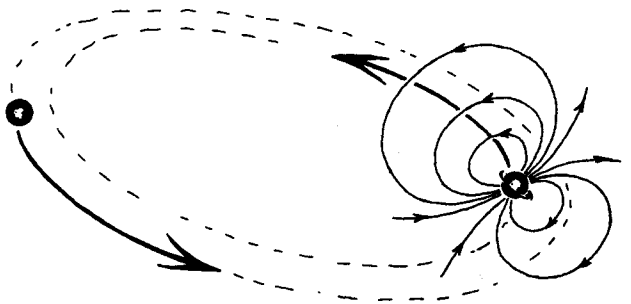


Рис. 4.9. Двойной пульсар PSR 1913+16. Две нейтронные звезды вращаются одна вокруг другой. Одна из звезд является пульсаром; ее магнитное поле чрезвычайно велико и способно захватывать заряженные частицы.

Результатом коллапса звезды является также огромное увеличение скорости ее вращения (как следствие сохранения кинетического момента). В случае нашего пульсара (диаметр около 20 км) скорость вращения, как мы уже говорили, составляет приблизительно 17 оборотов в секунду! В итоге магнитное поле пульсара также вращается со скоростью 17 оборотов в секунду, так как линии индукции внутри звезды остаются жестко связанными с телом звезды. Линии поля вне звезды увлекают

за собой заряженные частицы, однако на определенном расстоянии от звезды скорость, с которой этим частицам приходится перемещаться, приближается (причем вплотную) к скорости света. Оказавшись в такой ситуации, заряженные частицы принимаются интенсивно излучать в радиодиапазоне, и это чрезвычайно мощное излучение, подобно свету гигантского маяка, распространяется на огромное расстояние. Поскольку «маяк» вращается, Земли достигает лишь часть излучаемых им импульсов; астрономы наблюдают их в виде характерной для данного пульсара последовательности «радиощелчков» (рис. 4.10).

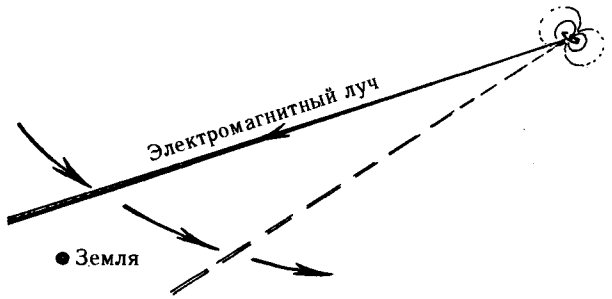



Рис. 4.10. Захваченные магнитным полем заряженные частицы вращаются вместе с пульсаром и испускают электромагнитный сигнал, который «накрывает» Землю 17 раз в секунду. Этот сигнал мы принимаем в виде последовательности коротких радиоимпульсов.

Скорости вращения пульсаров чрезвычайно стабильны — пульсары можно использовать как часы, причем точность этих часов будет сопоставима с точностью наиболее совершенных из существующих в данный момент на Земле часов (атомных) — а то и превзойдет ее. (Хорошие «пульсарные часы» спешат — или отстают — всего лишь на $10^{-12}$ с в год.) Если пульсар является частью системы двойной звезды (как, например, в случае с PSR 1913+16), то его орбитальное движение вокруг своего спутника можно точно регистрировать за счет *эффекта Доплера* — частота принимаемых на Земле щелчков несколько увеличивается, когда пульсар к нам приближается, и уменьшается, когда он удаляется.



В случае PSR 1913+16 астрономам удалось получить чрезвычайно подробную картину действительных взаимных орбит обеих звезд и убедиться в справедливости ряда различных предсказаний общей теории относительности Эйнштейна. Среди последних можно упомянуть эффект, называемый «смещением перигелия», — в конце XIX века астрономы обратили внимание на аномалии в орбитальном движении Меркурия вокруг Солнца, каковые аномалии Эйнштейн в 1916 году объяснил в рамках своей теории, что стало первым ее испытанием на прочность, — а также разного рода общерелятивистские «качания» и «вихляния», воздействующие на поведение осей вращения и тому подобных объектов. Поведение системы, состоящей из двух малых тел, движущихся друг вокруг друга по общей орбите, описывается в теории Эйнштейна очень четкой (детерминистской и вычислимой) моделью — движение тел в этом случае можно вычислить с высокой степенью точности, используя как сложные и тонкие методы аппроксимации, так и различные стандартные вычислительные методы. Некоторые необходимые для такого вычисления параметры нам точно не известны — например, массы и начальные скорости движения звезд, — впрочем, данных, извлеченных из сигналов пульсара, вполне достаточно для того, чтобы предсказать значения этих параметров с высокой точностью. Картина, получаемая в результате вычислений, замечательно согласуется, как в общем, так и в частностях, с информацией, содержащейся в принимаемых нами сигналах пульсара, что можно считать еще одним существенным подтверждением общей теории относительности.

Общая теория относительности предполагает существование еще одного эффекта, о котором я до сих пор не упоминал; между тем, он играет важную роль в динамике двойных пульсаров. Речь идет о *гравитационном излучении*. В предыдущем параграфе я отмечал, что гравитация существенным образом отличается от всех прочих физических взаимодействий. Тем не менее, в некоторых отношениях гравитация и электромагнетизм очень похожи. Среди прочего, электромагнитные поля обладают одним важным свойством: они способны существовать в волновой форме, распространяясь в пространстве в виде световых или радиоволн. Согласно классической теории Максвелла, источником таких волн становится любая система движущихся друг относительно друга заряженных частиц, взаимодействующих че-

рез посредство электромагнитных сил. Аналогичным образом, согласно классической общей теории относительности, источником гравитационных волн является любая система движущихся друг относительно друга гравитирующих тел — вследствие возникающих между ними гравитационных взаимодействий. При обычных обстоятельствах эти волны чрезвычайно слабы. Самым мощным источником гравитационного излучения в Солнечной системе является движение Юпитера вокруг Солнца, но при этом количества гравитационной энергии, испускаемой системой Солнце—Юпитер, едва хватит на то, чтобы зажечь сорокаваттную лампочку!

Однако при иных условиях — например, в системе двойного пульсара PSR 1913+16 — ситуация коренным образом меняется, и гравитационное излучение системы начинает играть весьма существенную роль. Теория Эйнштейна дает уверенный и детальный прогноз относительно природы гравитационного излучения подобных систем — в частности, предполагается, что система должна терять в процессе определенное количество энергии. В результате потери энергии должно происходить медленное сближение нейтронных звезд по спирали; соответственно, должен уменьшаться и период их обращения друг вокруг друга. Первыми двойной пульсар PSR 1913+16 наблюдали Джозеф Тейлор и Расселл Халс в 1974 году, с помощью гигантского радиотелескопа «Аресибо», расположенного в Пуэрто-Рико. Впоследствии Тейлор и его коллеги регулярно измеряли период обращения звезд этого пульсара и установили, что он уменьшается в точном соответствии с предсказанием общей теории относительности (см. рис. 4.11). За эту работу Тейлор и Халс получили в 1993 году Нобелевскую премию по физике. Наблюдение за системой PSR 1913+16 продолжается до сих пор, и чем больше данных мы накапливаем, тем больше подтверждений эйнштейновской теории получаем. В самом деле, если взять систему в целом и сравнить наблюдаемое ее поведение с поведением, рассчитанным по теории Эйнштейна (также взятой в целом), — начиная с ньютоновских расположений орбит, далее внося в эти орбиты поправки на стандартные эффекты общей теории относительности и завершая всю процедуру учетом эффекта потери энергии при гравитационном излучении, — то мы обнаружим, что теория полностью подтверждается, при этом погрешность составляет не более $10^{-14}$. Таким образом, можно смело утверждать, что эйнштейновская общая теория относительности является, в данном конкретном смысле, наиболее тщательно проверенной теорией из всех известных науке!



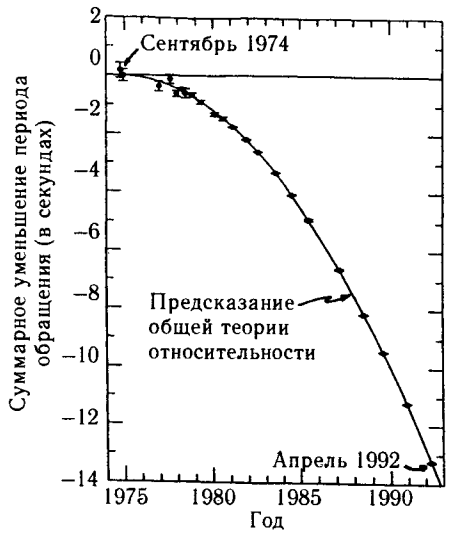



Рис. 4.11. Этот график (любезно предоставленный Дж. Тейлором) демонстрирует точное согласие наблюдаемого (на протяжении 20 лет) уменьшения периода взаимного обращения составляющих пульсар нейтронных звезд с расчетной потерей энергии системой при гравитационном излучении в соответствии с теорией Эйнштейна.

В описанном примере мы рассматриваем систему в высшей степени «чистую» — при ее расчете необходимо учитывать только эффекты общей теории относительности. Не нужно беспокоиться ни о сложностях, связанных с учетом внутреннего строения входящих в систему тел, ни о замедлении их движения под воздействием промежуточной среды или магнитных полей — все это не оказывает на динамику системы сколько-нибудь заметного влияния. Более того, мы имеем здесь дело лишь с двумя телами и их совокупным гравитационным полем, поэтому выполнить полное и точное вычисление их ожидаемого поведения — в рамках

теории, исчерпывающе описывающей все существенные аспекты этого самого поведения — нам вполне по силам. Возможно, на сегодняшний день, это один из наиболее выдающихся примеров совершенного согласия между расчетной теоретической моделью и экспериментально наблюдаемым поведением (для систем, состоящих из малого количества тел).

Даже если тел в физической системе значительно больше, модель поведения системы все равно можно рассчитать с той же точностью, воспользовавшись возможностями, предоставляемыми современными компьютерными технологиями. В частности, имеется очень подробная и полная модель движения всех планет Солнечной системы вместе с их наиболее значительными спутниками, построенная Ирвином Шапиро и его коллегами. Эту модель можно рассматривать как еще одно существенное подтверждение общей теории относительности. Здесь теория Эйнштейна также согласуется со всеми результатами наблюдений и прекрасно объясняет всевозможные малые отклонения от наблюдаемого движения, возникающие в моделях, использующих исключительно ньютоновский подход.

С помощью современных компьютеров можно выполнить расчеты и для систем, содержащих еще большее количество тел — порой порядка миллиона, — хотя такие расчеты, как правило (но не всегда), вынуждены целиком и полностью опираться на теорию Ньютона. Приходится прибегать к некоторым упрощающим допущениям — например, не рассчитывать воздействие буквально каждой частицы на все остальные, а как-то аппроксимировать воздействие всей совокупности частиц с помощью того или иного усреднения. Подобные методы вычислений широко распространены в астрофизике, где тщательно исследуются процессы формирования звезд и галактик, а также «догалактического» сгущения материи.

Впрочем, между предполагаемыми целями тех и других вычислений имеется существенная разница. В данном случае нас, конечно же, интересует отнюдь не *действительная* эволюция некоторой системы, но ее *типичная* эволюция. Как и в рассмотренном нами ранее случае хаотических систем, такой подход будет здесь, пожалуй, наиболее оправданным. С его помощью можно исследовать различные научные гипотезы о составе и первоначальном распределении материи во Вселенной, чтобы убедиться, насколько хорошо, в общем и целом, результаты описываемой в



этих гипотезах эволюции согласуются с тем, что мы наблюдаем на деле. При таких обстоятельствах никто и не ожидает получить соответствие в мельчайших деталях, но сравнить общую картину и различные статистические параметры модели и наблюдаемого феномена вполне возможно.

Крайний случай такого рода возникает, когда количество частиц настолько велико, что нет никакой надежды проследить эволюцию каждой из них в отдельности, — частицы в таких системах исследуются исключительно статистическими методами. Так, общепринятое математическое описание газа оперирует статистическими *ансамблями* различных возможных движений частиц, не размениваясь на частные движения каждой отдельной частицы. Температура, давление, энтропия и прочие подобные физические величины являются характеристиками как раз таких ансамблей, но эти же характеристики можно считать и частью вычислительной системы, в которой эволюционные свойства ансамблей рассматриваются со статистической точки зрения.

Помимо соответствующих динамических уравнений (Ньютона, Максвелла, Эйнштейна или кого угодно еще), исследователь таких систем должен взять на вооружение еще один физический принцип — *второй закон термодинамики*[5]. Нужен он, в сущности, для того, чтобы исключить из рассмотрения те начальные состояния движения отдельных частиц, что ведут к совершенно невероятным, хотя и возможным динамически, эволюциям. Применение второго закона позволяет гарантировать, что данная эволюция моделируемой системы действительно является «типичной», что мы не получим в результате наших усилий *атипичную* модель, не имеющую к решаемой задаче никакого практического отношения. С помощью второго закона можно довольно точно рассчитывать дальнейшую эволюцию систем, содержащих огромное количество частиц, отследить движение каждой из которых мы физически не в состоянии.

Зададим себе интересный — и весьма непростой — вопрос: почему, несмотря на то, что динамические уравнения Ньютона, Максвелла и Эйнштейна абсолютно симметричны во времени, упомянутые эволюции невозможно достоверно распространить в *прошлое*? Почему в реальном мире второй закон термодинамики в обратном направлении не работает? Причина имеет, очевидно, самое непосредственное отношение к весьма особым условиям, существовавшим в начале времени, — иначе говоря, к возник-

новению Вселенной в результате Большого Взрыва. (Подробное обсуждение гипотезы Большого Взрыва см. в НРК, глава 7.) Более того, эти начальные условия оказываются особыми ровно настолько, что благодаря им мы получаем еще один пример чрезвычайно высокой точности моделирования наблюдаемого физического поведения посредством четко сформулированных математических гипотез.

Что касается Большого Взрыва, то существенным элементом соответствующих гипотез является то, что на самых ранних его стадиях составляющая Вселенную материя находилась в состоянии *теплового равновесия*. Что же такое «тепловое равновесие»? Исследование состояний теплового равновесия — это крайность, противоположная точному моделированию движения небольшого количества объектов (предпринятому, например, в вышеописанном случае двойного пульсара). Здесь нас интересует исключительно «типичное поведение» в его чистейшем и наиболее наглядном виде. Состояние равновесия — это, вообще говоря, состояние системы, которая полностью «устоялась» и не намерена из этого своего состояния выходить, даже если ее слегка «потревожить». В случае систем с большим количеством частиц (или с большим количеством степеней свободы) — т. е. там, где рассматривается уже не движение каждой отдельной частицы, но усредненное поведение этих частиц и усредненные же параметры (например, температура и давление), — состоянием, в которое в конечном счете, согласно второму закону термодинамики (принцип максимума энтропии), приходит система, будет именно состояние *теплового* равновесия. Уточнение «теплового» в данном случае подразумевает, что речь идет о некотором усреднении разнонаправленного движения большого количества отдельных частиц, составляющих систему. Именно средние и составляют предмет исследования в термодинамике — т. е. поведение не индивидуальное, но типичное.

Строго говоря, из всего вышеизложенного следует, что когда речь заходит о термодинамическом состоянии системы или о тепловом равновесии, под этим вовсе не подразумевается какое-то индивидуальное состояние — скорее, имеется в виду некая совокупность, или ансамбль, состояний, которые на макроскопическом уровне представляются совершенно одинаковыми (а энтропия, если не вдаваться в детали, есть не что иное, как логарифм количества состояний в этом ансамбле). Если взять некоторое количество газа в состоянии равновесия и определить его давление, объем, а также количество и расположение молекул газа, то мы получим весьма характерное распределение вероятных скоростей частиц при тепловом равновесии (впервые это распределение было описано Максвеллом). При более тщательном анализе обнаруживается масштаб, в котором следует ожидать статистических флуктуаций от идеального состояния теплового равновесия, и здесь мы вступаем во владения более сложной науки, называемой *статистической механикой*, — науки о статистическом поведении материи.



Может показаться, что и в моделировании физического поведения посредством математических структур также нет ничего принципиально невычислимого. После выполнения соответствующих расчетов мы, как правило, приходим к хорошему согласию между вычисленным и наблюдаемым. Однако если рассматриваемая система хоть сколько-нибудь сложнее, нежели заполненное разреженным газом пространство или обширная совокупность гравитирующих тел, нам вряд ли удастся полностью избежать проблем, обусловленных *квантовомеханической* природой составляющей систему материи. Даже такой чистейший и наиболее тщательно исследованный образчик термодинамического поведения, как состояние теплового равновесия между веществом и излучением (так называемое «*абсолютно черное тело*»), нельзя исчерпывающе описать в классических терминах — необходимо учитывать и квантовые процессы, происходящие на фундаментальном уровне. Более того, у истоков всей квантовой теории лежит не что иное, как предпринятая Максом Планком в 1900 году попытка анализа излучения черного тела.

Как бы то ни было, предсказания физической теории (а ныне — квантовой теории) блестяще подтверждаются. Наблюдаемая экспериментально взаимосвязь между частотой и интенсивностью излучения на этой частоте весьма точно описывается предложенной Планком формулой. Хотя в рамках настоящего рассуждения нас, вообще говоря, интересует вычислительная природа *классической* теории, я не в силах устоять перед искушением привести пример наиболее совершенного (на сегодняшний день и насколько мне известно) согласия между данными наблюдений и результатами вычислений по формуле Планка. Этот пример можно также рассматривать как превосходное экспериментальное подтверждение стандартной модели

Большого Взрыва — в том, что имеет отношение к температурным условиям в новоиспеченной Вселенной в первые несколько минут ее существования. На рис. 4.12 маленькими прямоугольниками показаны экспериментальные значения интенсивности космического фонового излучения на различных частотах (полученные с помощью исследовательского спутника COBE[3]); непрерывная кривая построена в соответствии с формулой Планка, при этом за температуру фонового излучения взято значение 2,735 (±0,06) K (наилучшее эмпирическое значение). Точность совпадения кривых поражает воображение.


Рис. 4.12. Точное согласие между результатами наблюдений, полученными со спутника COBE, и теоретическими результатами в предположении «тепловой» природы излучения Большого Взрыва.

Приведенные выше примеры взяты из астрофизики — области, особое внимание в которой уделяется именно сравнению результатов громоздких вычислений с наблюдаемым поведением существующих в реальном мире систем. Прямые эксперименты в астрофизике невозможны, поэтому подтверждения теориям приходится искать путем сравнения рассчитанного (исходя из стандартных физических законов) поведения той или иной системы в той или иной предполагаемой ситуации с данными, полученными с помощью сложных наблюдательных процедур. (Наблюдения осуществляются с поверхности Земли, с аэростатов или других

[3]Cosmic Background Explorer (англ.) — *букв.* «Исследователь космического фонового излучения». — *Прим. перев.*



летательных аппаратов, размещенных в верхних слоях атмосферы, с ракет или искусственных спутников; при этом наряду с обычными оптическими телескопами применяются и самые разнообразные детекторы прочих сигналов.) Все эти вычисления, впрочем, не имеют непосредственного отношения к цели наших поисков, и я упомянул о них, главным образом, как о замечательно наглядных примерах того, насколько продуктивным инструментом исследования природы могут оказаться полные и точные вычисления, насколько хорошо вычислительные процедуры способны в действительности подражать природе. Нам же стоит уделить более пристальное внимание исследованиям биологических систем, так как именно в поведении биологических систем (а точнее — согласно выводам, к которым мы пришли в первой части, — в поведении осознающего себя мозга) следует искать возможные и необходимые проявления невычислимой физической активности.

Нет никаких сомнений в том, что вычислительные модели играют весьма важную роль в моделировании биологических систем, однако сами эти системы очевидно гораздо более сложны, чем те, с которыми имеет дело астрофизика, — соответственно, более сложной оказывается и задача построения действительно надежной модели биологической системы. Количество систем, достаточно «чистых» для того, чтобы получить при моделировании сколько-нибудь «приличную» точность, очень невелико. Мы в состоянии построить достаточно эффективные модели сравнительно простых систем — таких, например, как кровоток в сосудах различных типов или, скажем, передача сигналов по нервным волокнам (хотя в последнем случае возникают некоторые сомнения относительно того, допустимо ли рассматривать данную систему в рамках исключительно классической физики, поскольку важную роль здесь играют, наряду с физическими, и химические процессы).

Химические процессы напрямую обусловлены квантовыми эффектами, поэтому при исследовании поведения, связанного с химической активностью, мы, строго говоря, выходим за рамки классической физики. Несмотря на это, очень часто подобные «квантово обусловленные» процессы рассматриваются с позиций существенно классических. И хотя формально такой подход корректным не является, в большинстве случаев мы интуитивно предполагаем, что всевозможные тонкие квантовые эффек-

ты (помимо тех, что «официально» учитываются стандартными правилами и законами химии, классической физики и геометрии) серьезной роли здесь не играют. С другой стороны, мне думается, что при всей разумности и даже беспроигрышности такого предположения в отношении моделирования многих биологических систем (сюда, пожалуй, можно включить и распространение нервных импульсов) все же несколько рискованно делать общие выводы о более сложных биологических процессах, опираясь лишь на их якобы полностью классическую природу, особенно если речь заходит о таких сложнейших системах, как, например, человеческий мозг. Если мы намерены прийти к сколько-нибудь общим заключениям о теоретической возможности достоверной вычислительной модели мозга, нам необходимо прежде как-то разобраться с «загадками» квантовой теории.

Именно этим мы и займемся в двух последующих главах — по крайней мере, попытаемся по мере возможности. Там, где, как мне представляется, разобраться в причудах квантовой теории невозможно в принципе, я покажу, каким образом следует модифицировать саму теорию с тем, чтобы привести ее в вид, более соответствующий нашим представлениям о правдоподобной картине мира.

## Примечания

1. См., напр., [81], с. 49.
2. Одно из таких соотношений — «первый закон термодинамики»: $dE = TdS - pdV$. Буквами $E$, $T$, $S$, $p$ и $V$ здесь обозначены, соответственно, энергия, температура, энтропия, давление и объем газа.
3. См., напр., [81].
4. [333]; см. также [265], с. 428.
5. Весьма живописное, но не очень детальное изложение сути второго закона термодинамики имеется в НРК (глава 6). Интересующихся подробностями отсылаю к [69], а тех, кто не боится трудностей, — к [288].

# 5

# СТРУКТУРА КВАНТОВОГО МИРА

## 5.1. Квантовая теория: головоломки и парадоксы

Квантовая теория дает нам превосходное описание физической реальности на микроскопическом уровне, однако полна при этом тайн и загадок. Нет никакого сомнения: разобраться в том, как именно работает эта теория, чрезвычайно трудно; еще труднее отыскать какой-либо смысл в той «физической реальности» (или нереальности), которая, как утверждает квантовая теория, и составляет основу нашего мира. На первый, неискушенный, взгляд может показаться, что эта теория способствует формированию мировоззрения, которое многие (включая и меня) находят в высшей степени неудовлетворительным. В лучшем случае, буквально понимая все положения и определения теории, мы получаем, мягко говоря, очень странную картину мира. В худшем — столь же буквально воспринимая заявления некоторых из наиболее знаменитых приверженцев квантовой теории, никакой картины мира мы не получаем вовсе, а та, что была, рассыпается на глазах.

Я думаю, все те загадки, что ставит перед нами квантовая теория, можно четко разделить на два совершенно различных класса. Одни я называю *загадками-головоломками*, или **Z-загадками** (от слова *puzzle*[1]). К этому классу я отношу те квантовые истины об окружающем нас мире, которые действительно способны кого угодно привести в замешательство и заставляют изрядно поломать над собой голову — и в то же время находят непосредственное экспериментальное подтверждение. Сюда же можно включить и те общие предсказания квантовой теории, которые не подтверждены экспериментально, но — ввиду

[1] Головоломка (англ.). — *Прим. перев.*

уже подтвержденного — очень похожи на правду. Среди наиболее поразительных **Z**-загадок упомяну те, что известны под общим названием *феномены Эйнштейна – Подольского – Розена* (или ЭПР-феномены; подробнее о них мы поговорим позднее, см. §§ 5.4, 6.5). Второй класс составляют квантовые загадки, которые я называю *загадками-парадоксами*, или **X**-*загадками* (от слова *paradox*[2]). Согласно квантовому формализму, эти утверждения о мире вроде бы должны быть истинными, однако они настолько невероятны и парадоксальны, что мы просто не можем в них поверить, не можем признать их «действительно» истинными. Именно эти загадки и не дают нам принять предлагаемый формализм всерьез, препятствуют образованию на рассматриваемом уровне сколько-нибудь достоверной картины мира. Самая знаменитая **X**-загадка — парадокс *шрёдингеровой кошки*, в рамках которого, по всей видимости, утверждается, что макроскопические объекты (например, кошки) способны существовать в двух совершенно различных состояниях одновременно (этакое подвешенное состояние, в котором кошка и «жива», и «мертва» сразу). К подобным парадоксам мы еще вернемся в § 6.6 (см. также § 6.9, рис. 6.3, и НРК, с. 290–293).

Нередко утверждают, что все трудности, которые возникают у наших современников с восприятием квантовой теории, происходят исключительно от того, что мы чересчур крепко цепляемся за наши старые физические концепции. С каждым же последующим поколением люди будут «вживаться» в квантовые таинства все глубже, и в конце концов, после достаточного количества сменившихся поколений, смогут без какого-либо напряжения принять их все скопом — как **Z**-загадки, так и **X**-загадки. Этот взгляд представляется мне фундаментально ошибочным.

Я полагаю, что к **Z**-загадкам мы, возможно, и в самом деле сможем со временем привыкнуть и даже счесть их вполне естественными, однако с **X**-загадками такой номер *не* пройдет. По моему глубокому убеждению, **X**-загадки заведомо неприемлемы с философской точки зрения, а возникновение их объясняется только тем, что квантовая теория не является полной теорией — или, скорее, не является вполне точной на том уровне феноменов, на котором начинают проявляться **X**-загадки. В совершенной квантовой теории ни одной **X**-загадки в списке квантовых тайн не

[2]Парадокс (англ.). — *Прим. перев.*



останется (а *крест* в их названии оказался символичен — им и перечеркнем). Иначе говоря, свыкаться нам предстоит лишь с **Z**-загадками.

Учитывая вышесказанное, мы имеем полное право поинтересоваться, где же проходит граница между **Z**-загадками и **X**-загадками. Одни физики утверждают, что квантовых загадок, которые следовало бы в этом смысле классифицировать как **X**-загадки, попросту нет, — *все* странные и на первый взгляд парадоксальные утверждения, в которые нам предлагает поверить квантовый формализм, действительно истинны и описывают реальный мир, нужно только правильным образом на этот самый мир посмотреть. (Если такие люди хотят избежать обвинений в отсутствии логики и всерьез воспринимают возможность описания физической реальности в терминах «квантовых состояний», то они должны также верить и во «множественность миров» в той или иной форме (см. § 6.2). Согласно этой концепции, шрёдингеровы мертвая и живая кошки обитают в различных «параллельных» вселенных. Вы видите кошку, и тут же в каждой из двух вселенных возникает по вашей копии, один из вас глядит на живую кошку, а другой — на мертвую.) Другие физики устремляются к противоположной крайности. По их мнению, я слишком благодушно настроен по отношению к квантовому формализму, раз полагаю, что всем этим необъяснимым ЭПР-феноменам (о которых, напоминаю, мы еще поговорим) и впрямь найдется в будущем экспериментальное подтверждение. Я никоим образом не настаиваю, что все должны непременно разделять мое мнение о том, где именно надлежит проводить границу между **Z**- и **X**-загадками. Мой выбор определяется предположениями, согласующимися с точкой зрения, которую я представлю в следующей главе, в § 6.12.

Вряд ли уместно будет приводить на этих страницах исчерпывающее объяснение природы квантовой теории. Поэтому в настоящей главе я ограничусь относительно кратким (но в достаточной мере полным) описанием некоторых необходимых нам аспектов теории, особое внимание уделив при этом природе **Z**-загадок. В следующей главе я расскажу, почему я полагаю, что наличие **X**-загадок делает современную квантовую теорию неполной, невзирая на все те поразительные экспериментальные подтверждения, которыми она на сегодняшний день может похвастаться. Читателям, желающим познакомиться с квантовой

теорией поближе, я рекомендую обратиться к НРК (глава 6) или к более специальной литературе — например, [94], или [70].

Далее (глава 6, § 6.12) я представлю одну новую идею относительно уровня, на котором имеет смысл предпринимать попытки усовершенствования квантовой теории (думаю, следует предупредить читателя, что идея эта существенно отличается от той, что была предложена в НРК, хотя мотивы остались почти теми же). В § 7.10 (и в § 7.8) я приведу некоторые предварительные причины, позволяющие предположить, что подобные попытки вполне могут быть связаны с невычислимостью в том общем смысле, который нас так интересует. Что касается *стандартной* квантовой теории, то невычислимой она является лишь постольку, поскольку в измерительной процедуре здесь наличествуют случайные элементы. Случайные же элементы, как я особо подчеркивал в первой части (§§ 3.18, 3.19), не способны сами по себе обусловить ту невычислимость, которая нам потребуется в конечном итоге для понимания процессов мышления.

Рассмотрим для начала некоторые из наиболее поразительных **Z**-загадок квантовой теории на примере двух весьма показательных и мозгодробительных головоломок.

## 5.2. Задача Элитцура – Вайдмана об испытании бомб

Вообразим себе бомбу, в носовой части которой закреплен детонатор, настолько чувствительный, что при малейшем давлении на него бомба взрывается. Для срабатывания такого детонатора достаточно одного-единственного фотона видимого света, хотя в некоторых случаях детонатор заклинивает, и бомба взорваться не может — бомбу с неисправным детонатором мы будем называть «холостой». Предположим, что детонатор снабжен зеркальцем, подвижно закрепленным на носу бомбы таким образом, что при отражении зеркальцем одного фотона (видимого света) оно смещается и приводит в движение ударный механизм, в результате чего бомба взрывается — за исключением, разумеется, тех случаев, когда бомба оказывается холостой, т. е. когда чувствительный механизм детонатора заклинивает. Поскольку все упомянутые устройства работают по классическим законам, мы должны также предположить, что после того, как бомба собрана,



выяснить, не заклинило ли ее детонатор, невозможно без того, чтобы этот самый детонатор так или иначе не потревожить — что непременно приведет к немедленному взрыву. (Необходимо ввести еще одно допущение: детонатор может заклинить только в процессе сборки, по завершении сборки детонатор либо исправен, либо нет; см. рис. 5.1.)

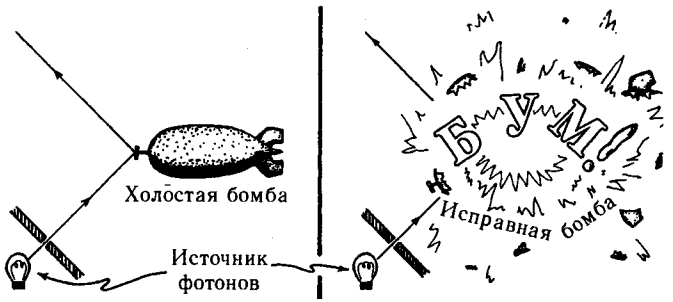



Рис. 5.1. Задача Элитцура – Вайдмана об испытании бомб. Сверхчувствительный детонатор бомбы срабатывает от соприкосновения с одним-единственным оптическим фотоном — может, впрочем, и не сработать, если его заклинит, в каковом случае бомба считается холостой. Задача: найти гарантированно исправную бомбу при наличии большого количества бомб сомнительного качества.

Допустим, что таких бомб у нас огромное количество (денег мы здесь не считаем!), однако доля холостых среди них может оказаться чрезмерно высокой. Задача заключается в том, чтобы найти хотя бы одну бомбу, о которой можно было бы заранее с уверенностью сказать: «Вот эта точно сработает».

Эта задача (вместе с решением) была предложена Авшаломом Элитцуром и Львом Вайдманом [114]. Я не буду приводить решение прямо здесь, так как, возможно, кто-то из читателей, уже знакомых с квантовой теорией и с теми занимательными головоломками, которые я определил выше как **Z**-загадки, пожелает попробовать свои силы (интеллектуальные, разумеется) в отыскании этого самого решения. Достаточно будет сказать, что решение существует и даже, при неограниченном запасе бомб такого рода, не выходит за рамки современных технических воз-

можностей. Тех же, кто в квантовой теории пока не сведущ (либо просто не склонен тратить время на поиски решения), я прошу потерпеть еще некоторое время (или, если хотите, можете сразу заглянуть в §5.9). Всему свое время — сначала я попытаюсь объяснить некоторые фундаментальные квантовые идеи, а затем приведу решение.

На данном этапе рассуждения необходимо лишь отметить: одно то, что эта задача имеет-таки решение (квантовомеханическое), уже указывает на глубинное различие между квантовой и классической физикой. При классическом подходе выяснить, не заклинило ли детонатор бомбы, можно только посредством приложения к нему какого-либо *реального* физического усилия (при этом, если детонатор исправен, бомба взрывается, и эксперимент считается благополучно проваленным). В рамках квантовой теории возможны и иные варианты — например, физический эффект, являющийся результатом того, что к детонатору *могло* быть приложено усилие, в то время как в действительности ничего подобного *не произошло*. В этом, собственно, и состоит одна из наиболее любопытных особенностей квантовой теории: реальный физический эффект здесь вполне может являться результатом *контрфактуальных* (как говорят философы) действий, т. е. действий, которые могли произойти, хотя на деле и не произошли. При рассмотрении следующей **Z**-загадки мы убедимся, что контрфактуальность играет далеко не последнюю роль и в ситуациях иного рода.

## 5.3. Магические додекаэдры

В качестве предисловия к нашей второй **Z**-загадке позвольте мне рассказать вам небольшую историю, не лишенную, впрочем, некоторой головоломности[1]. Представьте себе, получил я не так давно по почте замечательно выполненный правильный додекаэдр (рис. 5.2). Отправитель — компания «Квинтэссенциальные Товары», предприятие с превосходной репутацией и штабквартирой на одной из планет далекого красного гиганта, известного нам под названием Бетельгейзе. Точно такой же додекаэдр они отослали и моему коллеге, который в настоящий момент проживает на планете, обращающейся вокруг альфы Центавра, что приблизительно в четырех световых годах отсюда. Мне также



стало известно, что его додекаэдр прибыл к нему примерно в то же время, что и мой ко мне. На каждой вершине обоих додекаэдров имеется по кнопке. Нам с коллегой предлагается нажимать кнопки на наших додекаэдрах — по одной за раз. Выбор кнопок, порядок и время их нажатия оставлены целиком и полностью на наше усмотрение. Иногда при нажатии кнопки ничего не происходит, в каковом случае нам следует перейти к следующей кнопке. Может, впрочем, произойти следующее событие: зазвенит звонок, за чем последует впечатляющий фейерверк, сопровождающийся полным разрушением данного конкретного додекаэдра.

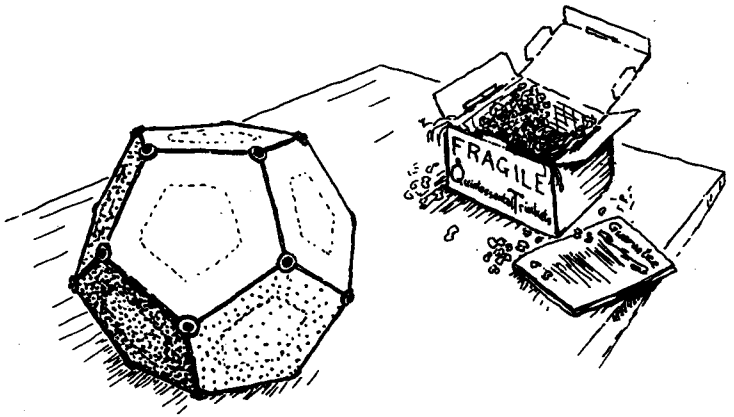



Рис. 5.2. Магический додекаэдр. У моего коллеги из системы альфы Центавра есть точно такой же. На каждой из вершин имеется кнопка. Результатом нажатия на какую-либо из кнопок *может* стать звонок и впечатляющий фейерверк. (FRAGILE = НЕ БРОСАТЬ; Quintessential Trinkets = Квинтэссенциальные Товары; Guarantee = Гарантии)

В коробку вместе с каждым додекаэдром был вложен перечень свойств, гарантированно присущих как моему додекаэдру, так и додекаэдру моего коллеги. Прежде всего нам следует очень тщательно расположить наши додекаэдры в пространстве таким образом, чтобы они были сориентированы совершенно одина-

ково. «Квинтэссенциальные Товары» предоставили и подробные инструкции, описывающие, как именно нужно располагать наши додекаэдры относительно, скажем, центров Туманности Андромеды и галактики М-87 и т. д. Самое главное здесь — добиться полной идентичности в ориентации наших двух додекаэдров. Перечень гарантированных свойств достаточно обширен, но нам понадобятся лишь некоторые из них, да и те довольно просты.

Следует учесть, что компания «Квинтэссенциальные Товары» производит подобные вещи уже очень долго — скажем, сотню миллионов лет или около того, — и никто никогда не смог уличить ее в том, что гарантированные ею свойства поставляемых устройств не соответствуют действительности. Эта надежность и составляет основу той безупречной репутации, которую компания поддерживает вот уже миллион столетий, поэтому мы можем быть совершенно уверены — если компания заявляет, что ее товар обладает тем или иным свойством, то так оно, безусловно, и есть. Более того, компания объявила, что выплатит некую ошеломительную ПРЕМИЮ любому, кто обнаружит-таки в гарантированных свойствах обман или ошибку, и никто пока за вознаграждением не обращался!

Нас с вами интересуют те из гарантированных свойств, которые касаются последовательности нажатия кнопок. Мы с коллегой независимо друг от друга выбираем одну из вершин своего додекаэдра. Такие вершины я буду называть ВЫБРАННЫМИ. Причем соответствующие кнопки мы *не нажимаем*. Вместо этого мы нажимаем по очереди (в любом порядке, как нам заблагорассудится) те три кнопки, что располагаются в вершинах, *соседних* с ВЫБРАННОЙ. Если при нажатии на одну из этих кнопок зазвенит звонок, то все операции с данным конкретным додекаэдром придется, разумеется, прекратить, однако он вполне может и не зазвенеть. Нам понадобятся следующие два свойства (см. рис. 5.3):

(а) если в качестве соответствующих ВЫБРАННЫХ вершин мы с коллегой вдруг выберем вершины диаметрально *противоположные*, то при одном из моих нажатий (на кнопки, соседние с ВЫБРАННОЙ вершиной) звонок может зазвенеть только в том случае, если он звенит при нажатии моим коллегой кнопки при диаметрально противоположной вершине, — независимо от порядка, в каком нам заблагорассудится упомянутые кнопки нажимать;

(б) если же в качестве соответствующих ВЫБРАННЫХ вершин мы с коллегой выберем *одинаковые* вершины (т. е. те, направления на которые из центров додекаэдров совпадают), звонок должен зазвенеть при нажатии, по крайней мере, на одну кнопку из наших общих шести.



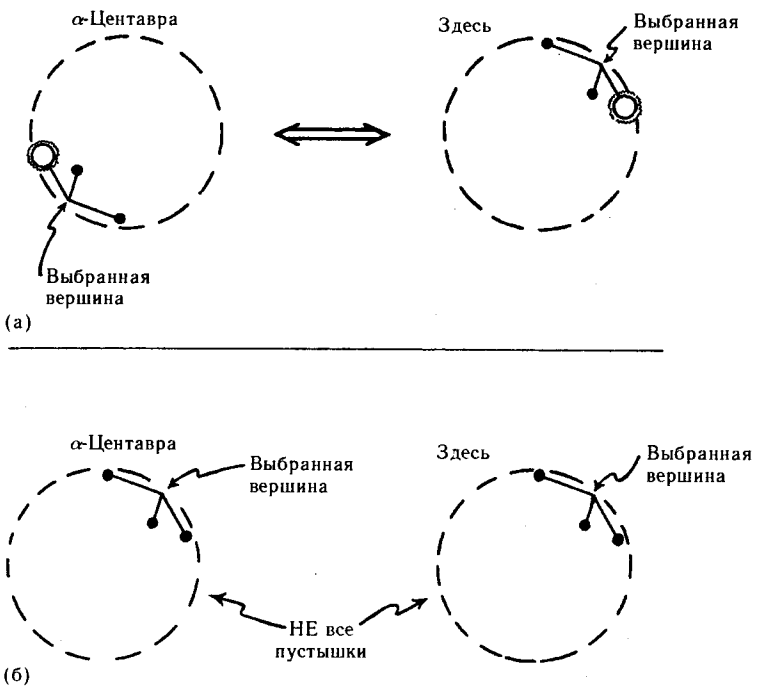



Рис. 5.3. Свойства додекаэдров, гарантируемые компанией «Квинтэссенциальные Товары». (а) Если мы с коллегой ВЫБИРАЕМ *противоположные* вершины додекаэдра, то звонок может зазвенеть только при нажатии диаметрально противоположных кнопок, независимо от порядка нажатия. (б) Если мы ВЫБИРАЕМ *одинаковые* вершины, то при нажатии какой-то из шести кнопок звонок непременно зазвенит.

Теперь я попробую сделать кое-какие выводы о правилах, которым должен подчиняться *мой* додекаэдр (независимо от того, что там происходит на альфе Центавра), на основании того простого факта, что «Квинтэссенциальные Товары» оказываются каким-то образом способны давать столь нерушимые гарантии, не имея ни малейшего представления о том, какие именно кнопки мне или моему коллеге придет в голову нажать. В качестве ключевого допущения предположим, что никакой дальнодействующей «связи» между моим додекаэдром и додекаэдром моего коллеги нет. Будем считать, что после того, как наши додекаэдры покинули «сборочный цех», они существуют раздельно и совершенно независимо друг от друга. Выводы следующие (рис. 5.4):

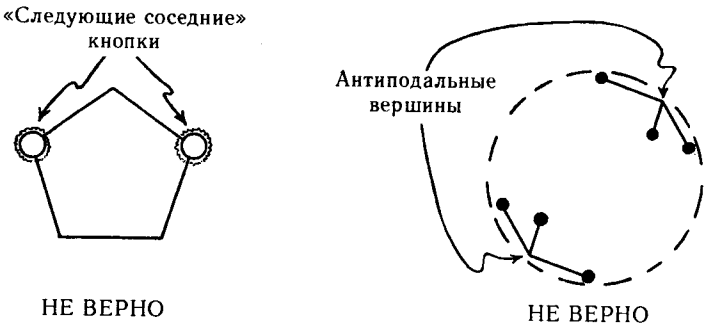



Рис. 5.4. Предположим, что наши додекаэдры представляют собой независимые (никак не связанные друг с другом) объекты. Тогда каждая кнопка на моем додекаэдре заведомо является либо звонком (БЕЛЫЕ кнопки), либо пустышкой (ЧЕРНЫЕ кнопки), при этом две соседние кнопки не могут обе быть БЕЛЫМИ, и никакой набор из шести кнопок при вершинах, соседних с двумя антиподальными вершинами, не может состоять из одних ЧЕРНЫХ кнопок.

(в) каждая из кнопок при вершинах моего додекаэдра заведомо является либо звонком (обозначим такие вершины БЕЛЫМ цветом), либо пустышкой (обозначим ЧЕРНЫМ), при этом ее «звонковость» никак не зависит от того, нажимаю я ее первой, второй или третьей из кнопок при вершинах, соседних с ВЫБРАННОЙ;



(г) две «следующие соседние» кнопки не могут обе быть звонками (т. е. БЕЛЫМИ кнопками);

(д) никакой набор из шести кнопок при вершинах, соседних с двумя антиподальными вершинами, не может состоять из одних пустышек (т. е. ЧЕРНЫХ кнопок)

(*Антиподальными* я здесь называю диаметрально противоположные вершины одного додекаэдра.)

Утверждение (в) мы выводим из того факта, что вполне *может* случиться так, что мой коллега выберет в качестве ВЫБРАННОЙ вершины вершину, диаметрально противоположную моей ВЫБРАННОЙ вершине; по крайней мере, «Квинтэссенциальным Товарам» неоткуда узнать заранее, что он ее не выберет (вот она, контрфактуальность!). Таким образом, если в результате какого-либо из моих нажатий зазвенит звонок, то кнопка при диаметрально противоположной вершине додекаэдра моего коллеги (*если* он нажмет ее первой из трех) тоже должна быть звонком. Так должно быть вне зависимости от того, в каком порядке я решил нажимать свои собственные три кнопки, а значит (исходя из допущения об отсутствии «связи» между додекаэдрами), мы с полной уверенностью можем сказать, что «Квинтэссенциальные Товары» изначально сделали кнопку при этой конкретной вершине звонком (в каком бы порядке я ни нажимал на свои кнопки), дабы избежать противоречия со свойством (а).

Аналогичным образом, из свойства (а) выводится утверждение (г). Предположим, что обе кнопки при двух следующих соседних вершинах являются звонками. Какую бы из этих кнопок я ни нажал первой, зазвенит звонок. Предположим теперь, что ВЫБРАННОЙ вершиной я назначил вершину, соседнюю им обеим. В этом случае порядок, в котором я нажимаю на свои кнопки, уже *имеет* значение, что противоречит свойству (а), если ВЫБРАННАЯ вершина додекаэдра моего коллеги противоположна ВЫБРАННОЙ вершине моего додекаэдра (а уж возможность такого совпадения «Квинтэссенциальные Товары» наверняка должны были учесть).

Наконец, учитывая то, что мы уже выяснили, мы легко выведем утверждение (д) из свойства (б). Предположим, что мы с коллегой выбираем в качестве ВЫБРАННЫХ *одинаково расположенные* вершины своих додекаэдров. Если ни одна из моих трех кнопок, соседних с ВЫБРАННОЙ вершиной, не является

звонком, то, согласно (б), звонком должна оказаться одна из трех соответствующих кнопок на додекаэдре моего коллеги. Из (а) следует, что кнопка моего додекаэдра, противоположная звонку на додекаэдре моего коллеги, также должна быть звонком. Получается (д).

А теперь, собственно, головоломка. Попробуйте окрасить каждую вершину додекаэдра в БЕЛЫЙ или ЧЕРНЫЙ цвет, строго следуя правилам (г) и (д). Очень скоро вы обнаружите, что как бы вы ни старались, ничего хорошего из этого не получается. В таком случае вот вам головоломка получше: *докажите*, что раскрасить вершины додекаэдра таким образом *невозможно*. Для того, чтобы дать всякому достаточно заинтригованному читателю шанс найти решение самостоятельно, я скромно помолчу до Приложения В (с. 467), где и приведу свое (боюсь, не очень изящное) доказательство того, что подобная раскраска действительно невозможна. Может быть, кому-то из читателей придет в голову что-нибудь более остроумное.

Неужели? Неужели, впервые за миллион столетий, «Квинтэссенциальные Товары» допустили наконец ошибку? Убедившись, что раскрасить вершины моего додекаэдра в соответствии с правилами (в), (г) и (д) *невозможно*, и ни на секунду не забывая о величине ожидающей нас ПРЕМИИ, мы, подпрыгивая на месте от нетерпения, ждем четыре (приблизительно) долгих года, по истечении которых приходит сообщение от моего коллеги, в котором подробно описано, какие он нажимал кнопки и когда, и не звенел ли звонок в его додекаэдре. Ознакомившись с сообщением, мы впадаем в уныние, а все наши надежды на ПРЕМИЮ тают как снег в жаркий день, потому что «Квинтэссенциальные Товары» снова подтвердили свою безупречную репутацию!

Рассуждения, приведенные в Приложении В (с. 467), однозначно демонстрируют, что в рамках любой классической модели просто-напросто *не существует* способа построить магические додекаэдры, обладающие теми свойствами, на которые «Квинтэссенциальные Товары» с такой легкостью выдают безусловную гарантию, — не существует, если исходить из допущения, что по окончании сборки два додекаэдра представляют собой абсолютно отдельные, никак не связанные друг с другом объекты. Ибо *никто не в состоянии* гарантировать наличие у двух додекаэдров требуемых свойств (а) и (б) без того, чтобы эти додекаэдры не были неким таинственным образом «связаны» друг с другом.



По крайней мере, в тот момент, когда мы начинаем нажимать на кнопки, эта «связь» должна наличествовать — кроме того, природа ее такова, что передача сигнала на расстояние около четырех световых лет осуществляется, по всей видимости, мгновенно. И все же «Квинтэссенциальные Товары» почему-то считают для себя возможным предоставлять такие гарантии — гарантии невозможного! — и никто до сих пор не смог уличить их в ошибке.

В чем же здесь подвох? Как «Квинтэссенциальные Товары» — или «КТ», эта аббревиатура хорошо известна многим их клиентам — умудряются проделывать такие фокусы? Вы говорите, вам всегда казалось, что КТ — это *квантовая теория*? Пусть так, не буду спорить. Так вот, что делают «КТ» — они просто берут и подвешивают в центре каждого из наших додекаэдров по одному атому, *спин* которого равен $\frac{3}{2}$, ни больше ни меньше. Эти два атома производятся на Бетельгейзе изначально вместе (общий спин пары равен 0), а затем аккуратно разделяются и помещаются в центры двух додекаэдров; общий спин связанной пары атомов при этом так и остается равным 0. (О том, что все это означает, мы поговорим в §5.10.) В результате, когда я нажимаю кнопку при одной из вершин своего додекаэдра (то же относится и к моему коллеге с его додекаэдром), производится некое измерение спина (неполное) в направлении от центра додекаэдра к данной конкретной вершине. Если результат измерения оказывается утвердительным, то звенит звонок, и через некоторое время додекаэдр рассыпается замечательным фейерверком. Более подробно о природе этого измерения я расскажу позднее (см. §5.18), а также покажу в §5.18 и Приложении В, почему правила (а) и (б) являются следствием из стандартных правил квантовой механики.

Замечательный вывод, который из всего этого следует, заключается в том, что допущение об отсутствии дальнодействующей «связи» между додекаэдрами к квантовой теории *неприменимо*! На пространственно-временно́й диаграмме (рис. 5.5) хорошо видно, что наши с коллегой нажатия на кнопки представляют собой *пространственноподобно разделенные* события (см. §4.4): согласно теории относительности, никакой обмен сигналами, передающими информацию о том, какие кнопки мы нажимаем или какие кнопки (на моей или на его стороне) окажутся в действительности звонками, между нами невозможен. Квантовая

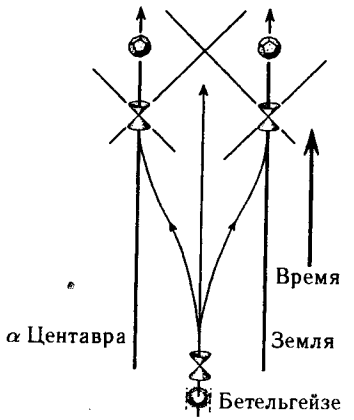



Рис. 5.5. Пространственно-временна́я диаграмма истории двух додекаэдров. Прибытие моего додекаэдра на Землю и прибытие додекаэдра моего коллеги на альфу Центавра — пространственноподобно разделенные события.

же теория, напротив, вполне допускает существование некоей «связи», соединяющей наши додекаэдры через пространственноподобно разделенные события. Вообще говоря, эту «связь» нельзя использовать для передачи непосредственно «пригодной к употреблению» информации, и в этом смысле никакого операционного конфликта между специальной теорией относительности и квантовой теорией нет. Имеет место лишь конфликт с *духом* специальной теории относительности — что, собственно, и является превосходной иллюстрацией одной из наиболее глубоких **Z**-загадок квантовой теории, феномена *квантовой нелокальности*. Два атома в центрах наших додекаэдров образуют *сцепленное состояние*, и, согласно правилам стандартной квантовой теории, их нельзя считать отдельными независимыми объектами.

## 5.4. Z-загадки ЭПР-типа: экспериментальный статус

Вышеприведенный эксперимент (мысленный, конечно же) относится к классу так называемых *ЭПР-измерений*, впервые



описанных в знаменитой статье Альберта Эйнштейна, Бориса Подольского и Натана Розена, опубликованной в 1935 году [113] (отсюда и название; подробнее об ЭПР-эффектах мы поговорим в § 5.17). В оригинальном варианте статьи речь шла, правда, не о спине, а об определенных комбинациях положения и импульса. Впоследствии Дэвид Бом включил в рассмотрение и спины — на примере пары частиц со спином $\frac{1}{2}$ (скажем, электронов), испускаемых из некоего источника в связанном состоянии со спином 0. На первый взгляд, из этих мысленных экспериментов следует, что измерение, произведенное в некоторой точке пространства на одной из частиц, составляющих квантовую пару, может мгновенно оказать некое весьма специфическое «воздействие» на другую частицу пары, причем эта другая частица может находиться на произвольно большом расстоянии от первой частицы. Впрочем, этим «воздействием» нельзя воспользоваться для передачи сколько-нибудь полезного послания от одной частицы к другой. В терминах квантовой теории говорят, что такие две частицы находятся в состоянии *сцепленности* друг с другом. Феномен квантовой сцепленности — истинная **Z**-загадка — был впервые отмечен Эрвином Шрёдингером [335].

Много позже Джон Белл в своей знаменитой теореме (1966, [21]) показал, что совместные вероятности различных измерений спина, производимых на любой паре сцепленных частиц, связаны определенными математическими соотношениями (известными ныне как неравенства Белла), с необходимостью следующими из того, что упомянутые частицы представляют собой отдельные независимые друг от друга сущности — каковыми они, собственно, и являются с точки зрения обыкновенной классической физики. Однако в квантовой теории эти соотношения могут нарушаться, причем весьма специфическим образом. Следовательно, открывается возможность для проведения реальных экспериментов с целью выяснить, наконец, действительно ли в реальных физических системах эти соотношения нарушаются, как утверждает квантовая теория, или же мы пока можем положиться на классическое представление, согласно которому пространственно разделенные объекты никоим образом не могут влиять друг на друга, а неравенства Белла с необходимостью выполняются. (Соответствующие примеры можно найти в НРК, с. 284, 301.)

В качестве наглядного примера того, чего *не* следует искать в понятии сцепленности, Джон Белл любил приводить *носки Бертлмана*. Бертлманом звали его коллегу, который неизменно появлялся на людях в носках разного цвета. Об этой причуде Бертлмана знали все. (Я сам встречал Бертлмана однажды, и на основании собственных наблюдений могу подтвердить: носки его действительно были разного цвета.) Таким образом, если кому-нибудь случалось заметить, что, скажем, левый носок Бертлмана сегодня, скажем, зеленого цвета, то этот кто-то мгновенно обретал знание о том, что правый носок Бертлмана *зеленым не является*. Тем не менее, вряд будет разумным сделать отсюда вывод, что левый носок Бертлмана способен неким таинственным образом оказывать мгновенное воздействие на правый носок Бертлмана. Эти два носка представляют собой независимые друг от друга объекты, и для того, чтобы «свойство отличия носков» всегда выполнялось, нет никакой нужды прибегать к услугам «Квинтэссенциальных Товаров». Такой эффект может быть легко организован силами самого Бертлмана, который возьмет себе за правило всегда, что бы ни случилось, надевать на ноги разные по цвету носки. Носки Бертлмана не вступают в противоречие с неравенствами Белла; никакой дальнодействующей «связи» между носками нет. Однако в случае магических додекаэдров производства «КТ» никакая «бертлмано-носочная» трактовка не в состоянии объяснить гарантированные свойства фигур. Именно в этом, собственно, и заключалась главная мысль предыдущего параграфа.

Через несколько лет после опубликования работы Белла был предложен[2] и впоследствии проведен[3] ряд натурных экспериментов. Кульминационным стал знаменитый парижский эксперимент Алена Аспекта (совместно с группой коллег, 1981), в рамках которого исследовалось поведение фотонов, образующих «сцепленную» пару (см. § 5.17): фотоны излучались в противоположных направлениях и улавливались детекторами, разнесенными на расстояние приблизительно 12 метров. Эксперимент блестяще оправдал возложенные на него надежды, установив физическую реальность **Z**-загадок ЭПР-типа (в полном соответствии с предсказанием стандартной квантовой теории) — и нарушив все, какие только можно, неравенства Белла (рис. 5.6).

Следует, впрочем, упомянуть, что несмотря на весьма хорошее согласие между результатами эксперимента Аспекта и




Рис. 5.6. ЭПР-эксперимент Алена Аспекта и его коллег. Пáры фотонов в сцепленном состоянии испускаются из источника. Решение о том, с какой стороны от источника измерять поляризацию фотона, принимается уже после того, как фотоны устремляются в разных направлениях, — исключая возможность передачи «сообщения» об этом решении от одного фотона другому.

предсказаниями квантовой теории, до сих пор есть еще физики, отнюдь не считающие, что эти результаты как-то подтверждают существование феномена квантовой нелокальности. Они указывают на то, что детекторы фотонов в эксперименте Аспекта (и в прочих подобных опытах) не обладали достаточной чувствительностью, вследствие чего бóльшую часть испущенных пар фотонов экспериментаторы в конечном итоге просто упустили. Последующая аргументация неизбежно приводит к следующему: если чувствительность детекторов повысить до некоторой пороговой степени, то пресловутое превосходное согласие между результатами наблюдений и предсказаниями квантовой теории рассеется как дым, немедленно восстановив в правах все те соотношения, которые, согласно Беллу, должны выполняться в любой локальной классической системе. Мне представляется крайне маловероятным, что то практически идеальное согласие квантовой теории и эксперимента, которое демонстрирует эксперимент Аспекта (см. рис. 5.7), окажется вдруг артефактом — более того, следствием *недостаточной чувствительности* детекторов. Еще менее правдоподобным выглядит предположение о том, что более совершенные детекторы каким-то образом это согласие ослабят — причем ослабят до такой степени, что можно будет говорить о справедливости в данном случае неравенств Белла[4].

Первоначально Белл получил соотношения между совместными *вероятностями* различных возможных событий (неравенства Белла). Для того чтобы оценить действительные вероят-

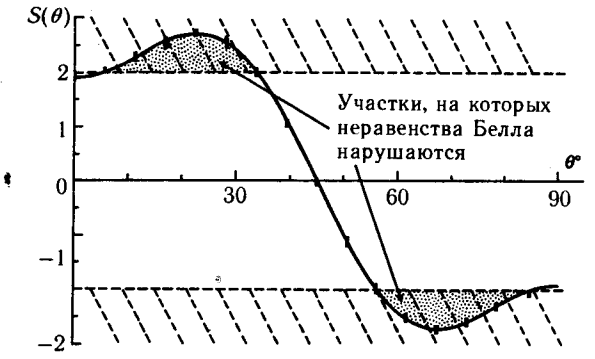



Рис. 5.7. Результаты эксперимента Аспекта очень хорошо согласуются с предсказаниями квантовой теории — и совершенно не вписываются в классические неравенства Белла. Неясно, каким образом более совершенные детекторы могут этому согласию помешать.

ности событий в рамках того или иного физического эксперимента, необходимо прежде накопить достаточный объем результатов наблюдений, а затем подвергнуть их соответствующему статистическому анализу. Не так давно был предложен ряд альтернативных проектов экспериментов (гипотетического характера), построенных исключительно на принципе «да/нет» и не нуждающихся в каком бы то ни было учете вероятностей. Первый из этих недавних проектов, разработанный в 1989 году Гринбергером, Хорном и Цайлингером [170], включает в себя измерение спина на частицах со спином $\frac{1}{2}$ в *трех* отдаленных друг от друга точках (скажем, на Земле, на альфе Центавра и на Сириусе — на случай, если этим проектом вдруг заинтересуются «Квинтэссенциальные Товары»). Ранее (в 1967 году) очень похожую идею выдвинули Кохен и Спекер [225], только они предполагали использовать частицы со спином 1 и чрезвычайно сложные геометрические конфигурации; да и сам Белл еще в 1966 году также работал над чем-то подобным, хотя и не столь конкретным [21]. (Эти ранние исследования, разумеется, не формулировались сразу в терминах ЭПР-феноменов; соответствующая переформулировка была предложена в 1983 году Хейвудом и Редхедом [197], см. также [358][(5)].) Приведенный выше пример с додекаэдрами хорош тем, что его геометрия весьма проста и легко представима визуально[(6)]. (Предлагались также эксперименты для изучения феноменов, эквивалентных уже упомянутым примерам **Z**-загадок, но иных физически; [394].)



## 5.5. Фундамент квантовой теории: исторический экскурс

Каковы же фундаментальные принципы квантовой механики? Прежде чем мы перейдем непосредственно к поискам ответа на этот вопрос, я хотел бы пригласить читателя на небольшую историческую экскурсию с целью проследить происхождение двух важнейших математических ингредиентов современной квантовой теории. При этом выяснятся совершенно замечательные (и малоизвестные широкой публике) вещи: во-первых, оба этих ингредиента появились, причем независимо друг от друга, еще в XVI веке, а во-вторых, придумал их *один и тот же* человек!

Человек этот, Джероламо Кардано (рис. 5.8), родился 24 сентября 1501 года в итальянском городе Павия, стал, помимо прочего, лучшим и известнейшим врачом своего времени и умер 20 сентября 1576 года в Риме. Несмотря на то, что его жизнь представляет собой один сплошной скандал (начиная с того, что союз его родителей не был освящен церковью, и заканчивая арестом и заключением в тюрьму уже самого Кардано на закате его жизни), он был человеком выдающегося ума и личных качеств, о чем, к сожалению, сегодня мало кому известно. Надеюсь, читатель простит меня, если я ненадолго отвлекусь от собственно квантовой механики и коротко расскажу об этом неординарном человеке.

В самом деле, в квантовой механике он совершенно неизвестен — зато его *имя* (все лучше, чем ничего) хорошо знакомо *автомеханикам. Карданным валом* называется универсальное устройство, соединяющее коробку передач автомобиля с его задними колесами и обеспечивающее гибкость, необходимую для поглощения переменного вертикального движения подрессоренной задней оси. Прототип этого изобретения Кардано

Рис. 5.8. Джероламо Кардано (1501–1576). Выдающийся врач, изобретатель, игрок, писатель и математик. Первооткрыватель комплексных чисел и теории вероятности — фундаментальных составляющих современной квантовой теории.

создал приблизительно в 1545 году, а в 1548 уже смог встроить его в шасси кареты, предназначенной для императора Карла V, что весьма скрасило тому путешествия по разбитым ухабистым дорогам. Кардано изобрел и многие другие полезные вещи — например, кодовый замóк, аналогичный тем, что используются в современных сейфах. Как врач, Кардано достиг широчайшей известности, среди его пациентов были короли и принцы. Он совершил множество открытий в медицине и написал немало книг на медицинские и другие темы. По всей видимости, именно Кардано первым указал, что такие венерические болезни, как сифилис и гонорея, представляют собой *разные* болезни и требуют, соответственно, *различного* лечения. Он же первым предложил лечить больных туберкулезом «санаторно» — на 300 лет раньше



Джорджа Боддингтона, который в 1830 году, в сущности, «переоткрыл» уже известное. В 1552 году Кардано вылечил Джона Гамильтона, архиепископа Шотландского, страдавшего астмой в тяжелой форме, — и оказал тем самым серьезное влияние на историю Британии.

Какое же отношение все эти впечатляющие достижения имеют к квантовой теории? Совершенно никакого, разве что демонстрируют широту ума человека, которому мы фактически обязаны открытием двух наиболее фундаментальных составляющих этой самой теории, причем открытия эти никак одно с другим не связаны. Кардано был выдающимся врачом и выдающимся изобретателем, однако этими областями деятельности он не ограничивался — он был еще и выдающимся математиком.

Первая из упомянутых составляющих — теория вероятностей. Как известно, квантовая теория является теорией скорее вероятностной, нежели детерминистской. Сами ее правила фундаментально обусловлены вероятностными законами. В 1524 году Кардано написал свою «Книгу об азартных играх» (*«Liber de Ludo Aleae»*), где заложил основы математической теории вероятностей. Описанные в книге законы Кардано сформулировал несколькими годами ранее и не преминул ими воспользоваться. Применение свежеоткрытых законов на практике (а вот и *выдающийся* игрок!) принесло ему достаточно денег для того, чтобы заплатить за обучение в медицинской школе в Павии. По всей видимости, Кардано с самых юных лет знал, что зарабатывать деньги *шулерством* — занятие весьма рискованное, поскольку именно в результате подобной деятельности был убит бывший муж его матери. Джероламо же обнаружил, что, используя открытые им законы, управляющие самим случаем, выигрывать можно вполне честно.

Вторая фундаментальная составляющая квантовой теории, открытая Кардано, — понятие *комплексного числа*. Комплексным называется число вида

$$a + ib,$$

где под $i$ понимается квадратный корень из минус единицы,

$$i = \sqrt{-1},$$

а $a$ и $b$ суть обычные вещественные числа (т. е. числа, которые можно представить в виде десятичных дробей). Сегодня мы назы-

ваем число *a вещественной* частью комплексного числа $a + ib$, а число $b$ — его *мнимой* частью. На эти странные числа Кардано наткнулся, пытаясь отыскать способ решения общего кубического уравнения. Кубическими называются уравнения вида

$$Ax^3 + Bx^2 + Cx + D = 0,$$

где $A$, $B$, $C$ и $D$ — некоторые заданные вещественные числа, а уравнение следует решать относительно $x$. В 1545 году Кардано опубликовал трактат под названием «*Ars magna*»[3], где и привел первый полный анализ решения таких уравнений.

С публикацией этого решения связана пренеприятнейшая история. Еще в 1539 году учитель математики Николо Фонтана, более известный по прозвищу Тарталья (что в переводе с итальянского означает «заика»), отыскал общее решение для некоторого широкого класса кубических уравнений. Тогда же Кардано подослал к нему одного своего приятеля, чтобы тот выведал у Тартальи, как выглядит это решение. Тарталья, однако, не пожелал о нем говорить, вследствие чего Кардано засел за работу и вскоре обнаружил искомое решение самостоятельно, опубликовав результат в 1540 году в своей книге «Практическая арифметика и простые измерения». Более того, Кардано удалось распространить свое решение на *все* возможные случаи; позднее Кардано описал этот общий аналитический метод решения в «*Ars magna*». В обеих книгах Кардано указывал на первенство Тартальи в отыскании решения для того класса случаев, где это решение применимо, однако в «*Ars magna*» он допустил ошибку, утверждая, что Тарталья дал ему разрешение на публикацию. Узнав об этом, Тарталья пришел в ярость и заявил, что он сам однажды рассказал Кардано (будучи у него в доме по какому-то делу) о своем решении, взяв с хозяина клятву, что тот никому и ни при каких обстоятельствах это решение не откроет. Как бы то ни было, Кардано оказался в непростой ситуации: публикуя свое решение, обобщающее ранее полученное решение Тартальи, он тем самым неизбежно раскрывал «тайну» этого частного случая. Единственным выходом, по всей видимости, было бы полное замалчивание уже полученных результатов и прекращение каких бы то ни было исследований в этой области — и вряд ли Кардано пошел бы на такое. Тарталья, однако, затаил на Кардано обиду и выжидал вплоть до 1570 года. Именно тогда, воспользовавшись тем, что репутация Кардано оказалась серьезно подмочена в силу других скандальных обстоятельств, Тарталья и нанес завершающий удар, приведший в конечном итоге к унижению и смерти Кардано. В тесном сотрудничестве с Инквизицией Тарталья собрал огромную коллекцию всевозможных улик против Кардано и лично организовал его арест и заключение под стражу. Освободили Кардано только в 1571 году, после того, как в Рим прибыл особый посланник от архиепископа Шотландского (которого, как мы помним, Кардано вылечил от астмы) с прошением об освобождении узника — «ученого, пекущегося лишь о сохранении и исцелении тел, дабы души Господни проживали в них весь отпущенный им срок».

[3] «Великое искусство» (лат.) — *Прим. перев.*



Вышеупомянутые «скандальные обстоятельства» включают в себя, в частности, суд над старшим сыном Кардано, Джованни Баттистой, по обвинению в убийстве. На суде Джероламо, рискнув своей репутацией, выступил с поручительством за сына. Это не принесло им обоим ничего хорошего, поскольку Джованни был-таки виновен — он убил жену (женился он, впрочем, не по своей воле), пытаясь прикрыть еще одно совершенное им же убийство. По всей видимости, убийство жены Джованни совершил по наущению и при содействии своего младшего брата Альдо (еще больший, как выясняется, негодяй: тогда же он предал Джованни, а позднее выдал собственного отца Инквизиции; наградой Альдо стало назначение его палачом Инквизиции в Болонье). Не способствовала восстановлению репутации Кардано и его дочь, которая умерла от сифилиса, приобретенного благодаря ее профессиональной деятельности — проституции.

Интересное упражнение в исторической психологии — попытаться понять, как же так вышло, что Джероламо Кардано, любящий, судя по всему, отец, преданный жене и детям, и вообще честный и чуткий человек, не лишенный высоких устремлений, воспитал столь недостойное потомство. Несомненно, от семейных забот его часто отвлекали другие интересы, многочисленные и требующие немалого времени. Несомненно, его более чем годичное (когда ему пришлось ехать в Шотландию для лечения архиепископа, хотя в первоначальной договоренности речь шла лишь о встрече в Париже) отсутствие дома после смерти жены очень неблагоприятно сказалось на детях. Несомненно также,

что в смерти жены непосредственно повинна убежденность Кардано в том, что ему самому звезды предсказали смерть в 1546 году, — чем ближе к этому сроку, тем больше погружался Кардано в лихорадочные исследования и запись еще не записанного, совершенно позабыв не только о детях, но и о жене, что и свело ее (а не его) в могилу к концу того самого года.

Сегодня Кардано известен гораздо меньше, чем он того заслуживает, и истоки этого забвения, как я подозреваю, кроются в его злосчастной судьбе и безнадежно запятнанной (совместными стараниями его детей, Инквизиции и — в особенности — Тартальи) репутации. В моей же личной «табели о рангах» он безоговорочно принадлежит к величайшим фигурам эпохи Возрождения. Несмотря на то, что Джероламо рос в бедности, на формирование его личности очень большое влияние оказала царившая в доме атмосфера стремления к знаниям. Его отец, Фацио Кардано, был увлечен геометрией; Джероламо вспоминал, как однажды, когда он был еще ребенком, отец взял его с собой в гости к Леонардо да Винчи и как взрослые засиделись за полночь, обсуждая какие-то геометрические задачи.

Что же касается опубликования Кардано раннего результата Тартальи и некорректного, мягко говоря, утверждения, что последний эту публикацию разрешил, то, думаю, большего уважения все же заслуживает желание сделать свое открытие достоянием общественности, нежели стремление утаить новые знания. Разумеется, Тарталью тоже можно понять — от сохранения открытий в тайне зависел, до некоторой степени, его достаток (особенно если учесть, что Тарталья являлся завсегдатаем публичных математических состязаний), однако именно трактат Кардано, включающий решение Тартальи в качестве частного случая, оказал серьезное и долговременное влияние на развитие математической науки. Более того, раз уж мы затронули вопрос первенства, то оно, судя по всему, принадлежит и вовсе третьему ученому — Сципионе дель Ферро, преподававшему в Болонском университете вплоть до своей смерти в 1526 году. Во всяком случае, в записях дель Ферро имеется то решение, которое позднее заново открыл Тарталья, хотя остается неясным, понимал ли дель Ферро, каким образом это решение можно модифицировать для описания случаев, рассмотренных Кардано в «*Ars magna*»; отсутствуют также какие бы то ни было свидетельства



в пользу того, что дель Ферро добрался до концепции комплексных чисел.

Для того чтобы понять, в чем заключается фундаментальность вклада Кардано, рассмотрим решение кубического уравнения более подробно. Воспользовавшись подстановкой $x \rightarrowtail x+a$, нетрудно свести общее кубическое уравнение к виду

$$x^3 = px + q,$$

где $p$ и $q$ — вещественные числа. С такой подстановкой математики XVI века были прекрасно знакомы. Однако если вспомнить о том, что числа, которые мы сегодня называем *отрицательными*, в те времена далеко не все считали «настоящими» числами, то можно предположить, что во избежание появления в окончательном уравнении отрицательных чисел, получаемые в результате уравнения имели несколько иной вид — в зависимости от знака при $p$ и $q$ (например, $x^3 + p'x = q$ или $x^3 + q' = px$). Чтобы не усложнять рассуждения без необходимости, я буду в дальнейшем придерживаться современного способа записи.

Решения вышеприведенного кубического уравнения можно представить графически. Для этого построим кривые $y = x^3$ и $y = px + q$ и отметим точки их пересечения. Координаты $x$ этих точек и будут искомыми решениями уравнения. Обратите внимание на рис. 5.9: функция $y = x^3$ представлена в виде кривой, а для прямой $y = px + q$ показаны несколько возможных вариантов. (Мне неизвестно, использовали ли Кардано или Тарталья такое графическое представление, хотя это вполне возможно. Здесь я использую его исключительно для удобства рассмотрения различных возможных случаев.) Те случаи, для которых годилось решение Тартальи, соответствуют в наших обозначениях прямым с отрицательным (или нулевым) $p$. В этих случаях прямая «опускается» слева направо, типичный пример — прямая $P$ на рис. 5.9. Отметим, что в таких случаях всегда существует только одна точка пересечения прямой и кривой, т. е. кубическое уравнение имеет лишь одно решение. В современных обозначениях мы можем записать решение Тартальи следующим образом:

$$x = \sqrt[3]{\left(w - \frac{1}{2}q\right)} - \sqrt[3]{\left(w + \frac{1}{2}q\right)},$$

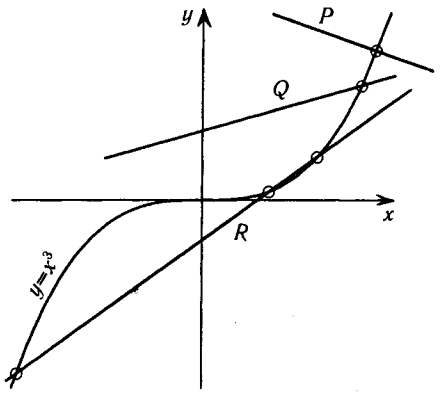



Рис. 5.9. Решения кубического уравнения $x^3 = px + q$ могут быть получены графически в виде точек пересечения прямой $y = px + q$ и кубической кривой $y = x^3$. Случай Тартальи охватывает прямые с $p \leqslant 0$ (на графике представлены убывающей прямой $P$), Кардано же описал и случаи с $p > 0$ (прямые $Q$ и $R$). *Casus irreducibilis* — случай с *тремя* точками пересечения (прямая $R$). В этом случае при записи решения возникает нужда в комплексных числах.

где

$$w = \sqrt{\left(\frac{1}{2}q\right)^2 + \left(\frac{1}{3}p'\right)^3}.$$

Через $p'$ мы здесь обозначаем $-p$; сделано это для того, чтобы все входящие в выражение величины оставались неотрицательными (число $q$ также выбирается положительным).

Обобщение Кардано этой процедуры учитывает также случаи $p > 0$ и позволяет записать решения для этих случаев (при положительном $p$ и отрицательном $q$; впрочем, знак при $q$ погоды не делает). Соответствующие прямые «поднимаются» слева направо (обозначены на рисунке буквами $Q$ и $R$). Мы видим, что при некотором заданном значении $p$ (т. е. при заданном угле наклона) и достаточно большом (т. е. таком, чтобы прямая пересекала ось $y$ в точке, расположенной достаточно высоко) $q'$ (иначе говоря, $-q$)



снова существует одно-единственное решение. Выражение Кардано для этого решения имеет вид (в современных обозначениях)

$$x = \sqrt[3]{\left(\frac{1}{2}q' + w\right)} - \sqrt[3]{\left(\frac{1}{2}q' - w\right)},$$

где

$$w = \sqrt{\left(\frac{1}{2}q'\right)^2 + \left(\frac{1}{3}p\right)^3}.$$

Вооружившись современными обозначениями и современной же концепцией отрицательного числа (а также учитывая тот факт, что кубический корень отрицательного числа равен отрицательному кубическому корню того же, но положительного числа), мы легко убеждаемся, что выражение Кардано, в сущности, идентично выражению Тартальи. Однако в случае Кардано в том же, казалось бы, выражении появляется нечто принципиально новое. Теперь при достаточно малом $q'$ прямая может пересечь кривую в *трех* точках, т. е. у исходного уравнения окажется три решения (при $p > 0$ два из них отрицательны). Случай этот — так называемый *casus irreducibilis*[4] — возникает, когда $\left(\frac{1}{2}q'\right)^2 < \left(\frac{1}{3}p\right)^3$; нетрудно видеть, что $w$ оказывается при этом *квадратным корнем из отрицательного числа*. Таким образом, числа $\frac{1}{2}q' + w$ и $\frac{1}{2}q' - w$ под знаком кубического корня в выражении Кардано являются не чем иным, как *комплексными числами*; сумма же этих двух кубических корней, если мы хотим получить решение уравнения, должна быть вещественным числом.

Это таинственное обстоятельство не избежало внимания Кардано, и позднее в «*Ars magna*» он отдельно обратился к вопросу, поставленному появлением комплексных чисел в решении уравнения, на примере задачи об отыскании двух чисел, произведение которых равно 40, а сумма равна 10. Эту задачу он решил (причем решил правильно), получив в качестве ответа два комплексных числа:

$$5 + \sqrt{-15} \quad \text{и} \quad 5 - \sqrt{-15}.$$

В графическом представлении задача сводится к отысканию точек пересечения кривой $xy = 40$ и прямой $x + y = 10$

[4] Неприводимый случай (лат.). — *Прим. перев.*

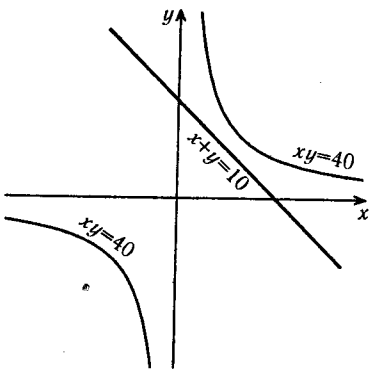



Рис. 5.10. Задача Кардано об отыскании двух чисел, произведение которых равно 40, а сумма равна 10, может быть представлена графически как отыскание точек пересечения кривой $xy = 40$ и прямой $x + y = 10$. При этом становится очевидным, что в вещественных числах эта задача решения не имеет.

(см. рис. 5.10). Отметим, что построенные на рисунке кривая и прямая нигде не пересекаются (в вещественных числах), что вполне согласуется с тем фактом, что для записи решения задачи требуются комплексные числа. Кардано эти новые числа в восторг отнюдь не приводили; он жаловался, что работа с ними «мучительна для разума». Тем не менее, изучая кубические уравнения, он вынужден был признать необходимость рассмотрения таких чисел.

Следует отметить, что необходимость в комплексных числах при записи решения кубического уравнения (представленного графически на рис. 5.9) обусловлена причинами, значительно более загадочными, нежели появление таких чисел в задаче, изображенной на рис. 5.10 (задача эта, в сущности, эквивалентна задаче отыскания корней квадратного уравнения $x^2 - 10x + 40 = 0$). В последнем случае вполне очевидно, что без привлечения комплексных чисел задача не имеет решения вовсе, и ничто не мешает нам объявить введение таких чисел безосновательной выдумкой, затеянной исключительно ради того, чтобы снабдить хоть каким-то «решением» уравнение, в действительности решений не имеющее. Эта позиция, однако, не объясняет, что происходит в случае кубического уравнения. Здесь (*casus irreducibilis* или прямая $R$ на рис. 5.9) уравнение действительно имеет три *вещественных* решения, отрицать существование которых невозможно, однако для того, чтобы выразить любое из этих решений даже в иррациональных числах (т. е. в квадратных и кубических корнях, как в данном случае), нам приходится забираться в таинственные дебри комплексных чисел, хотя окончательный результат и принадлежит миру чисел вещественных.



Похоже, что до Кардано никто в эти таинственные дебри не углублялся и не задумывался над тем, каким образом из них «произрастает» наш собственный «вещественный» мир. (Снаружи заглядывали — например, Герон Александрийский и Диофант Александрийский в первом и, соответственно, в третьем веках нашей эры, судя по некоторым свидетельствам, размышляли над идеей существования у отрицательного числа чего-то вроде «квадратного корня», однако ни один из них не набрался храбрости объединить такие «числа» с числами вещественными и прийти таким образом к понятию *комплексного* числа; не разглядели они и глубинной связи между своими «псевдочислами» и вещественными решениями уравнений.) Возможно, именно удивительное сочетание в одном человеке двух личностей — мистика и рационально мыслящего ученого — позволило Кардано уловить эти первые проблески того, что развилось позднее в одну из мощнейших математических концепций. В последующие годы, благодаря трудам Бомбелли, Коутса, Эйлера, Весселя, Арганда, Гаусса, Коши, Вейерштрасса, Римана, Леви, Льюи и многих других, теория комплексных чисел разрослась вглубь и вширь и занимает сегодня заслуженное место среди наиболее изящных и универсально применимых математических конструкций. Однако лишь с появлением в первой четверти двадцатого века квантовой теории мы осознали, какую странную и всепронизывающую роль играют комплексные числа в самóй фундаментальной структуре того физического мира, в котором мы живем, — не знали мы прежде и том, насколько тесна связь между комплексными числами и *вероятностями*. Даже у Кардано не возникло (да и не могло возникнуть) ни малейшего подозрения о существовании таинственной глубинной связи между двумя величайшими его вкладами в математику — связи, которая образует самый фундамент материальной Вселенной на тончайшем из ее уровней.

## 5.6. Основные правила квантовой теории

Что же это за связь? Что объединяет комплексные числа и теорию вероятностей, имея результатом неоспоримо превосходное описание работы тончайших внутренних механизмов нашего мира? Грубо говоря, законы комплексного исчисления справедливы на очень тонком подуровне феноменов, тогда как вероятности играют свою роль на узком мостике, что соединяет тот тонкий подуровень с хорошо знакомым нам уровнем обыденного восприятия, — от такого «объяснения», разумеется, проку немного; для сколько-нибудь реального понимания нам понадобится нечто более существенное.

Рассмотрим для начала роль комплексных чисел. В силу самого их определения их очень сложно принять в качестве инструмента для описания действительной физической реальности. Наибольшая сложность заключается в том, что им, на первый взгляд, просто нет места на уровне тех феноменов, что мы способны непосредственно воспринимать, на уровне, где действуют классические законы Ньютона, Максвелла и Эйнштейна. Таким образом, для того, чтобы наглядно представить себе, как именно работает квантовая теория, необходимо (хотя бы предварительно) учесть, что физические процессы происходят на двух четко разделенных уровнях: *квантовом* подуровне, где как раз и играют свою странную роль комплексные числа, и *классическом* уровне привычных макроскопических физических законов. На квантовом уровне комплексные числа выглядят вполне естественно — однако вся эта естественность напрочь пропадает, случись им забрести на уровень классический. Я вовсе не хочу сказать, что между уровнем, на котором действуют квантовые законы, и уровнем классически воспринимаемых феноменов непременно должно наличествовать физическое разделение; давайте просто вообразим (пока), что такое разделение существует — это поможет понять смысл процедур, реально применяемых в квантовой теории. Вопрос о существовании такого физического разделения в *действительности* очень глубок, и мы попытаемся на него ответить несколько позднее.

Где же *начинается* квантовый уровень? Надо думать, квантовым называется уровень тех физических объектов, которые «достаточно малы» — например, молекулы, атомы, элементарные частицы. Впрочем, на физические расстояния это требование



«малости» распространяется далеко не всегда. Эффекты квантового уровня могут возникать и на огромном удалении. Вспомним о четырех световых годах, разделяющих два додекаэдра в моей истории в § 5.3, или о двенадцати метрах, разделяющих фотоны во вполне реальном эксперименте Аспекта (§ 5.4). Иначе говоря, квантовый уровень определяется не малым физическим размером, но чем-то более тонким, причем на данном этапе этой «формулировкой» лучше и ограничиться. Можно также приблизительно считать квантовым уровень, где мы рассматриваем очень малые изменения в энергии. Более подробно мы обсудим этот вопрос в § 6.12.

Классическим же мы называем уровень, который мы, как правило, воспринимаем непосредственно. Здесь действуют законы классической физики, оперирующие вещественными числами, здесь имеют смысл самые обычные описания — например, те, что задают положение, скорость движения и форму футбольного мяча. Существует ли какая-либо реальная физическая граница между квантовым уровнем и уровнем классическим? Вопрос этот, как я только что отметил, очень глубок и тесно связан с трактовкой **X**-загадок, или квантовых парадоксов (см. § 5.1). Поиск ответа мы отложим до лучших времен, а пока, просто из соображений удобства, будем рассматривать квантовый уровень отдельно от классического.

Какую фундаментальную роль играют комплексные числа на квантовом уровне? Возьмем для примера отдельную частицу — скажем, электрон. В классической картине мира электрон может занимать либо положение **A**, либо какое-нибудь другое положение **B**. Однако в квантовомеханическом описании перед тем же электроном открываются гораздо более широкие возможности. Он не только может занимать то или иное из указанных положений, он может находиться и в любом из ряда возможных состояний, занимая при этом (в некотором строгом смысле) *оба* положения одновременно! Обозначим через $|\mathbf{A}\rangle$ состояние, в котором электрон занимает положение **A**, а через $|\mathbf{B}\rangle$ — состояние, в котором электрон занимает положение **B**.[5] Тогда, согласно квантовой

[5] Из соображений удобства я использую здесь предложенную Дираком стандартную систему обозначений для квантовых состояний (в данном случае, скобку «кет»). Читатели, незнакомые с квантовомеханическими обозначениями, могут пока не обращать на эти скобки внимания.

Поль Дирак был одним из наиболее выдающихся физиков двадцатого столетия.

теории, электрону доступны следующие возможные состояния:

$$w|\mathbf{A}\rangle + z|\mathbf{B}\rangle,$$

причем фигурирующие здесь весовые коэффициенты $w$ и $z$ представлены *комплексными числами* (и по крайней мере одно из них должно быть отлично от нуля).

Что это означает? Если бы весовые коэффициенты были неотрицательными *вещественными* числами, то можно было предположить, что записанная комбинация представляет собой, в некотором смысле, взвешенное вероятностное ожидание положения электрона, где $w$ и $z$ символизируют относительные вероятности нахождения электрона в положении, соответственно, **A** и **B**. Тогда отношение $w : z$ даст отношение вероятности нахождения электрона в точке **A** к вероятности нахождения электрона в точке **B**. Таким образом, если этими двумя и исчерпываются доступные электрону положения, то мы получаем ожидание $w/(w+z)$ для электрона в точке **A** и ожидание $z/(w+z)$ для электрона в точке **B**. При $w = 0$ электрон определенно находится в точке **B**; при $z = 0$ ищите его в точке **A**, больше ему деться некуда. Если состояние электрона записывается как $|\mathbf{A}\rangle + |\mathbf{B}\rangle$, это означает, что электрон может с равной вероятностью оказаться как в положении **A**, так и в положении **B**.

Однако числа $w$ и $z$ — *комплексные*, так что вышеприведенная интерпретация не имеет никакого смысла. Отношения квантовых весовых коэффициентов $w$ и $z$ *не являются* отношениями вероятностей. Это невозможно хотя бы потому, что вероятности всегда выражаются *вещественными* числами. Несмотря на широко распространенное мнение о вероятностной природе квантового мира, на квантовом уровне *не* действует карданова теория *вероятностей*. А вот его таинственная теория *комплексных чисел* пришлась здесь как нельзя более кстати — именно она лежит в основе математически точного и абсолютно *безвероятностного* описания процессов, протекающих на квантовом уровне.

Среди его достижений — общая формулировка законов квантовой теории, а также ее релятивистское обобщение, включающее в себя знаменитое «уравнение Дирака» для электрона. Дирак обладал удивительной способностью «чуять» истину — свои уравнения он оценивал в значительной степени по их *эстетическим* качествам!



Пользуясь привычным и понятным языком, невозможно объяснить, что «означает» фраза «в данный момент времени электрон находится в состоянии суперпозиции двух положений с комплексными весовыми коэффициентами $w$ и $z$». На настоящем этапе нам придется просто принять все это как должное; именно такими описаниями мы и вынуждены довольствоваться при рассмотрении квантовых систем. Такие суперпозиции, как сообщают естествоиспытатели, играют важную роль в действительной конструкции нашего микромира. Квантовый мир *на самом деле* ведет себя именно таким необычным и непостижимым образом, а нам повезло набрести на этот простой *факт*. А от фактов никуда не уйти — имеющиеся в нашем распоряжении описания, в соответствии с которыми эволюционирует микромир, действительно являются не только математически точными, но и, более того, *целиком и полностью детерминированными*!

## 5.7. Унитарная эволюция U

Таким детерминированным описанием является, например, *унитарная эволюция* (обозначим ее буквой **U**). Эта эволюция описывается точными математическими уравнениями, однако нам не так уж важно знать, как именно эти уравнения выглядят. Нам понадобятся лишь некоторые из свойств эволюции **U**. В так называемом «шрёдингеровом представлении» **U** задается уравнением Шрёдингера, которое характеризует скорость изменения *квантового состояния* (или *волновой функции*) во времени. Это квантовое состояние (обычно обозначаемое греческой буквой $\psi$, или так: $|\psi\rangle$) представляет собой полную взвешенную сумму (с комплексными весовыми коэффициентами) всех возможных альтернатив, доступных данной квантовой системе. Таким образом, для приведенного выше примера с двумя альтернативными положениями электрона квантовое состояние $|\psi\rangle$ записывается в виде следующей комбинации комплексных чисел:

$$|\psi\rangle = w|\mathbf{A}\rangle + z|\mathbf{B}\rangle,$$

где $w$ и $z$ — комплексные числа (причем хотя бы одно из них не равно нулю). Комбинацию $w|\mathbf{A}\rangle + z|\mathbf{B}\rangle$ мы называем *линейной суперпозицией* состояний $|\mathbf{A}\rangle$ и $|\mathbf{B}\rangle$. Величина $|\psi\rangle$ (равно как и $|\mathbf{A}\rangle$ или $|\mathbf{B}\rangle$) часто называется *вектором состояния*. Кванто-

вые состояния (или векторы состояния) могут записываться и в более общем виде — например, так:

$$|\psi\rangle = u|\mathbf{A}\rangle + v|\mathbf{B}\rangle + w|\mathbf{C}\rangle + \ldots + z|\mathbf{F}\rangle,$$

где $u$, $v$, ..., $z$ — комплексные числа (причем хотя бы одно из них не равно нулю), а $|\mathbf{A}\rangle$, $|\mathbf{B}\rangle$, ..., $|\mathbf{F}\rangle$ символизируют различные возможные положения, которые может занимать частица (или какое-либо иное возможное свойство частицы — например, ее спиновое состояние; см. § 5.10). Обобщая далее, можно допустить выражение волновой функции или вектора состояния в виде *бесконечной* суммы (поскольку число положений, которые может занимать точечная частица, бесконечно велико); впрочем, подобные случаи нас пока не занимают.

Здесь *необходимо* упомянуть об одной технической особенности квантового формализма. Дело в том, что значимыми являются только *отношения* комплексных весовых факторов. Подробнее об этом я расскажу позднее. А пока мы просто отметим, что для любого отдельно взятого вектора состояния $|\psi\rangle$ верно следующее: любое комплексное кратное $u|\psi\rangle$ (где $u \neq 0$) описывает то же самое *физическое* состояние, что и $|\psi\rangle$. Таким образом, например, физические состояния $uw|\mathbf{A}\rangle + uz|\mathbf{B}\rangle$ и $w|\mathbf{A}\rangle + z|\mathbf{B}\rangle$ совершенно идентичны. Соответственно, физический смысл имеет отношение $w : z$, но не отдельные числа $w$ и $z$.

Наиболее фундаментальным свойством уравнения Шрёдингера (а значит, и эволюции **U**) является его *линейность*. Иначе говоря, если у нас есть два состояния (скажем, $|\psi\rangle$ и $|\phi\rangle$) и уравнение Шрёдингера, согласно которому по прошествии времени $t$ состояния $|\psi\rangle$ и $|\phi\rangle$ эволюционируют в новые состояния, соответственно, $|\psi'\rangle$ и $|\phi'\rangle$, то любая линейная суперпозиция $w|\psi\rangle + z|\phi\rangle$ за то же время $t$ неминуемо эволюционирует в суперпозицию $w|\psi'\rangle + z|\phi'\rangle$. Для обозначения эволюции за время $t$ воспользуемся символом $\rightsquigarrow$. Тогда линейность подразумевает следующее: если

$$|\psi\rangle \rightsquigarrow |\psi'\rangle \qquad \text{и} \qquad |\phi\rangle \rightsquigarrow |\phi'\rangle,$$

то имеет место и эволюция

$$w|\psi\rangle + z|\phi\rangle \rightsquigarrow w|\psi'\rangle + |\phi'\rangle.$$

Это рассуждение применимо (разумеется) и к линейным суперпозициям трех и более индивидуальных квантовых состояний:



например, состояние $u|\chi\rangle + w|\psi\rangle + z|\phi\rangle$ эволюционирует за время $t$ в состояние $u|\chi'\rangle + w|\psi'\rangle + z|\phi'\rangle$, если каждое из состояний $|\chi\rangle$, $|\psi\rangle$ и $|\phi\rangle$ в отдельности эволюционирует за это же время, соответственно, в $|\chi'\rangle$, $|\psi'\rangle$ и $|\phi'\rangle$. Иными словами, эволюция всегда происходит так, словно каждый отдельно взятый компонент суперпозиции не «знает» о присутствии других. Можно сказать, что каждый отдельно взятый «мир», описываемый упомянутым компонентом, эволюционирует независимо от других, но всегда в соответствии с тем же уравнением Шрёдингера, что и другие. При этом комплексные весовые коэффициенты в суперпозиции, описывающей совокупное состояние, в процессе эволюции остаются неизменными.

Ввиду вышесказанного можно подумать, что суперпозиции и комплексные весовые коэффициенты не играют сколько-нибудь эффективной физической роли, поскольку эволюция отдельных состояний во времени происходит так, словно других состояний тут вовсе нет. Это заблуждение. Проиллюстрируем на примере, что может произойти с такой системой в реальности.

Рассмотрим случай падения света на полусеребрёное зеркало, т. е. на полупрозрачное зеркало, отражающее ровно половину падающего на него света и беспрепятственно пропускающее все остальное. По квантовой теории, свет образуют частицы, называемые *фотонами*. Вполне естественно будет предположить, что половина фотонов из падающего на полусеребрёное зеркало потока отражается от его поверхности, а половина проходит зеркало насквозь. Не тут-то было! Согласно все той же квантовой теории, при столкновении с поверхностью зеркала каждый *отдельный* фотон переходит в состояние *суперпозиции* отражения и пропускания. Если фотон находился до столкновения с зеркалом в состоянии $|\mathbf{A}\rangle$, то после столкновения состояние фотона эволюционирует (в соответствии с **U**) в состояние, которое можно записать в виде $|\mathbf{B}\rangle + i|\mathbf{C}\rangle$, где $|\mathbf{B}\rangle$ символизирует состояние, в котором фотон проникает сквозь зеркало, а $|\mathbf{C}\rangle$ — состояние, в котором фотон от зеркала отражается (см. рис. 5.11). Запишем эту эволюцию:

$$|\mathbf{A}\rangle \rightsquigarrow |\mathbf{B}\rangle + i|\mathbf{C}\rangle.$$

Коэффициент $i$ появляется здесь вследствие результирующего фазового сдвига на четверть длины волны[7], который возникает в таком зеркале между отраженным и прошедшим лучом света.

(Для большей точности мне следовало бы включить в выражение зависящий от времени коэффициент осцилляции и выполнить полную нормировку, однако в настоящем обсуждении никакой необходимости в такой точности нет. В приводимых описаниях я выделяю лишь существенные для нас аспекты происходящего. Несколько подробнее о коэффициенте осцилляции мы поговорим в § 5.11, а вопроса о нормировке коснемся в § 5.12. Более полное описание можно найти в любой стандартной работе по квантовой теории[8]; см. также НРК, с. 243–250.)

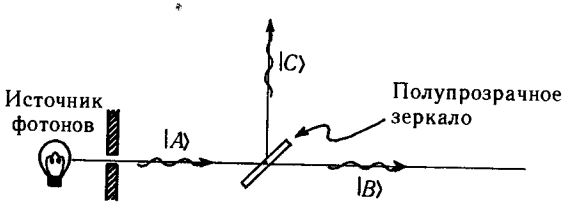



Рис. 5.11. Фотон в состоянии $|\mathbf{A}\rangle$ падает на полупрозрачное зеркало; в результате его состояние эволюционирует (согласно $\mathbf{U}$) в суперпозицию $|\mathbf{B}\rangle + i|\mathbf{C}\rangle$.

В рамках классической картины поведения частицы мы, разумеется, предположим, что состояния $|\mathbf{B}\rangle$ и $|\mathbf{C}\rangle$ представляют собой альтернативные варианты *возможного* поведения фотона. В квантовой же механике нам предлагается поверить, что фотон, находясь в такой чудесной комплексной суперпозиции, действительно совершает *оба указанных действия одновременно*. Чтобы убедиться в том, что здесь никоим образом не может идти речь о классических вероятностно-взвешенных альтернативах, разовьем наш пример еще немного и попытаемся снова свести вместе два частных состояния фотона (два фотонных луча). Для этого отразим сначала каждый луч от обычного, непрозрачного зеркала. В результате отражения[9] состояние $|\mathbf{B}\rangle$ фотона эволюционирует, согласно $\mathbf{U}$, в некоторое другое состояние, скажем, $i|\mathbf{D}\rangle$, тогда как состояние $|\mathbf{C}\rangle$ эволюционирует в $i|\mathbf{E}\rangle$:

$$|\mathbf{B}\rangle \rightsquigarrow i|\mathbf{D}\rangle \qquad \text{и} \qquad |\mathbf{C}\rangle \rightsquigarrow i|\mathbf{E}\rangle.$$

Таким образом, совокупное состояние $|\mathbf{B}\rangle + i|\mathbf{C}\rangle$ эволюционирует



по $\mathbf{U}$ следующим образом:

$$\begin{aligned} |\mathbf{B}\rangle + i|\mathbf{C}\rangle &\rightsquigarrow i|\mathbf{D}\rangle + i(i|\mathbf{E}\rangle) = \\ &= i|\mathbf{D}\rangle - |\mathbf{E}\rangle \end{aligned}$$

(поскольку $i^2 = -1$). Вообразим далее, что эти два луча сходятся на четвертом зеркале, на этот раз снова *полу*прозрачном (как показано на рис. 5.12; предполагается, что длины всех лучей одинаковы, благодаря чему коэффициент осцилляции, которым я по-прежнему пренебрегаю, не играет никакой роли и здесь). Состояние $|\mathbf{D}\rangle$ эволюционирует при этом в комбинацию $|\mathbf{G}\rangle + i|\mathbf{F}\rangle$, где $|\mathbf{G}\rangle$ представляет состояние прохождения, а $|\mathbf{F}\rangle$ — состояние отражения. Аналогичным образом, $|\mathbf{E}\rangle$ эволюционирует в $|\mathbf{F}\rangle + i|\mathbf{G}\rangle$, поскольку в этом случае $|\mathbf{F}\rangle$ символизирует состояние прохождения, а $|\mathbf{G}\rangle$ — состояние отражения:

$$|\mathbf{D}\rangle \overset{\rightsquigarrow}{\neq} |\mathbf{G}\rangle + i|\mathbf{F}\rangle \qquad \text{и} \qquad |\mathbf{E}\rangle \overset{\rightsquigarrow}{\neq} |\mathbf{F}\rangle + i|\mathbf{G}\rangle.$$

Нетрудно убедиться (ввиду линейности эволюции $\mathbf{U}$), что совокупное состояние $i|\mathbf{D}\rangle - |\mathbf{E}\rangle$ эволюционирует следующим образом:

$$\begin{aligned} i|\mathbf{D}\rangle - |\mathbf{E}\rangle &\rightsquigarrow i(|\mathbf{G}\rangle + i|\mathbf{F}\rangle) - (|\mathbf{F}\rangle + i|\mathbf{G}\rangle) = \\ &= i|\mathbf{G}\rangle - |\mathbf{F}\rangle - |\mathbf{F}\rangle - i|\mathbf{G}\rangle = \\ &= -2|\mathbf{F}\rangle. \end{aligned}$$

(Коэффициент $-2$ физического смысла не имеет, поскольку, как уже упоминалось выше, при умножении совокупного физического состояния системы — в данном случае, $|\mathbf{F}\rangle$ — на некоторое отличное от нуля комплексное число физическая ситуация остается прежней.) Таким образом, мы видим, что возможность $|\mathbf{G}\rangle$ оказывается для фотона *закрытой*: после слияния двух лучей в один открытой остается *единственно* возможность $|\mathbf{F}\rangle$. Этот любопытный результат обусловлен тем, что в физическом состоянии фотона в промежутке между его столкновениями с первым и последним зеркалом присутствуют *оба* луча *одновременно*. Мы говорим, что при этом происходит *интерференция* двух лучей. Как следствие, получается, что альтернативные «миры» фотона между упомянутыми столкновениями не отделены в действительности один от другого, но могут друг на друга влиять посредством этих самых феноменов интерференции.

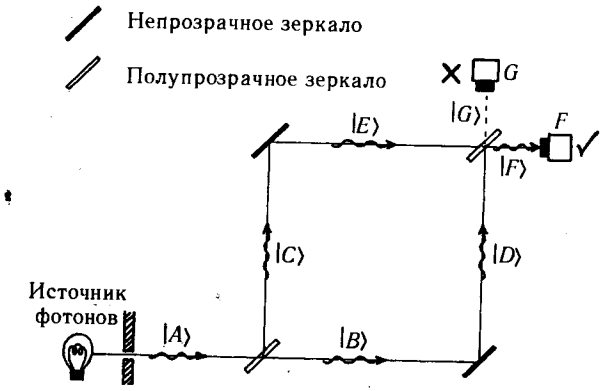



Рис. 5.12. Две составляющие состояния фотона сводятся вместе посредством двух непрозрачных зеркал; в точке слияния двух лучей установлено еще одно полупрозрачное зеркало. Лучи интерферируют таким образом, что результирующий луч приобретает состояние $|\mathbf{F}\rangle$, тогда как детектор в точке **G** фотона не регистрирует.

Важно помнить о том, что описанное свойство демонстрируют *единичные* фотоны. Следует понимать, что каждый отдельный фотон «пробует» оба открытых перед ним пути, оставаясь при этом все тем же *одним* фотоном. Он не расщепляется на два фотона на некоем промежуточном этапе, однако местоположение его определяется этаким странным комплексно-взвешенным *сосуществованием* альтернатив, что как раз и характерно для квантовой теории.

## 5.8. Редукция R вектора состояния

В рассмотренном выше примере суперпозиция состояний фотона переходит в конечном счете в одно-единственное состояние. Представим, что в точках, обозначенных на рис. 5.12 буквами **F** и **G**, размещены детекторы фотонов (фотоэлементы). Поскольку в данном конкретном примере фотон, миновав последнее зеркало, оказывается в состоянии $|\mathbf{F}\rangle$ (точнее, пропорциональном $|\mathbf{F}\rangle$), а состояние $|\mathbf{G}\rangle$ никакого участия в его дальнейшей судьбе не принимает, детектор в точке **F** зарегистрирует фотон, а детектор в точке **G** не зарегистрирует ничего.



Что произойдет в более общем случае — например, если мы попытаемся подать на эти детекторы суперпозицию состояний вроде $w|\mathbf{F}\rangle + z|\mathbf{G}\rangle$? Детекторы выполнят *измерение* с целью определить, находится фотон в состоянии $|\mathbf{F}\rangle$ или же в состоянии $|\mathbf{G}\rangle$. Квантовое измерение равносильно разглядыванию квантового события через увеличительное стекло и переводит событие с квантового на классический уровень. На квантовом уровне, при непрерывном воздействии **U**-эволюции, линейные суперпозиции сохраняются. Однако как только мы вытягиваем процесс на классический уровень, на котором события уже можно рассматривать как нечто *действительно* произошедшее, выясняется, что объекты больше не находятся в прежних странных комплексно-взвешенных комбинациях состояний. *Выясняется* (в нашем примере), что фотон регистрируется *либо* детектором в точке **F**, *либо* детектором в точке **G**, причем эти альтернативные варианты реализуются с определенной вероятностью. Квантовое состояние таинственным образом «перескакивает» от суперпозиции $w|\mathbf{F}\rangle + z|\mathbf{G}\rangle$ к состоянию «либо $|\mathbf{F}\rangle$, либо $|\mathbf{G}\rangle$». Такой «скачок» в описании состояния системы (от суперпозиции состояний квантового уровня к состоянию, при котором реализуется лишь одна из возможных альтернатив классического уровня) называется *редукцией вектора состояния*, или *коллапсом волновой функции*; эту операцию я буду обозначать буквой **R**. Вопрос о том, следует ли рассматривать операцию **R** как реальный физический процесс либо как некую иллюзию или аппроксимацию, чрезвычайно для наших целей важен, и мы к нему еще обязательно вернемся. Тот факт, что нам приходится (во всяком случае, в математических описаниях) отбрасывать эволюцию **U** и заменять ее совершенно отличной от нее процедурой **R**, есть фундаментальная **X**-загадка квантовой теории. На данном этапе, думаю, будет лучше, если мы не станем слишком углубляться в исследование этого парадокса, а будем (условно) рассматривать **R** как, в сущности, некий процесс, который *просто сопутствует* (в используемых нами математических описаниях, по крайней мере) процедуре «перемещения» события с квантового уровня на классический.

Как же вычисляются *вероятности* альтернативных результатов измерения на суперпозиции состояний? Для этого имеется одно весьма замечательное правило. Допустим, для измерения, определяющего окончательный выбор между альтернатив-

ными состояниями $|\mathbf{F}\rangle$ и $|\mathbf{G}\rangle$, как в приведенном выше примере, мы используем детекторы в точках, соответственно, **F** и **G**. Согласно упомянутому правилу, в случае суперпозиции состояний

$$w|\mathbf{F}\rangle + z|\mathbf{G}\rangle$$

отношение вероятности того, что фотон будет зарегистрирован детектором **F**, к вероятности того, что фотон будет зарегистрирован детектором **G**, равно

$$|w|^2 : |z|^2,$$

т. е. отношению *квадратов модулей* комплексных чисел $w$ и $z$. Квадрат модуля комплексного числа равен сумме квадратов его вещественной и мнимой частей; т. е. квадрат модуля числа

$$z = x + iy,$$

где $x$ и $y$ — вещественные числа, равен

$$\begin{aligned} |z|^2 &= x^2 + y^2 = \\ &= (x + iy)(x - iy) = \\ &= z\bar{z}. \end{aligned}$$

Число $\bar{z}$ $(= x - iy)$ называется *комплексным сопряженным* числа $z$; аналогичная операция проделывается и с $w$. (В вышеприведенном рассуждении я неявно подразумеваю, что состояния, обозначенные мною через $|\mathbf{F}\rangle$, $|\mathbf{G}\rangle$ и т. д., должным образом *нормированы*. Смысл этого термина я объясню позднее, см. § 5.12; строго говоря, нормировка необходима для того, чтобы выполнялось правило вероятностей в указанной форме.)

Именно здесь, и только здесь, на квантовую сцену выходят кардановы *вероятности*. Мы видим, что на квантовом уровне комплексные весовые коэффициенты *не* играют сами по себе роли относительных вероятностей (да и не могут этого делать, поскольку они комплексные), а вот вполне вещественные *квадраты модулей* этих комплексных коэффициентов такие роли играют. Более того, только теперь, после выполнения *измерений*, приобретают смысл понятия неопределенности и вероятности. Измерение квантового состояния происходит, в сущности, тогда, когда имеет место значительное «увеличение» некоторого физического процесса, вытягивающее его с квантового на классический уровень. В случае фотоэлемента регистрация квантового события — в виде приема фотона — вызывает в конечном счете



возмущение на классическом уровне, скажем, вполне отчетливый «щелчок». Вместо фотоэлемента мы могли бы использовать для регистрации фотона высокочувствительную фотографическую пластинку. В этом случае квантовое событие «прибытие фотона» вытягивается на классический уровень в виде хорошо различимой отметки на пластинке. В каждом из случаев измерительное устройство включает в себя некую неустойчиво уравновешенную систему — ничтожно малого квантового события оказывается достаточно, чтобы нарушить это равновесие и вызвать значительно больший по масштабу и наблюдаемый на классическом уровне эффект. Именно при этом переходе от квантового уровня к классическому комплексные числа Кардано возводятся в квадрат и становятся вероятностями Кардано!

Посмотрим, как можно применить это правило к конкретной ситуации. Предположим, что вместо зеркала в правом нижнем углу установлен фотоэлемент; тогда падающий на него фотон находится в состоянии

$$|\mathbf{B}\rangle + i|\mathbf{C}\rangle,$$

где состояние $|\mathbf{B}\rangle$ означает, что фотон регистрируется фотоэлементом, тогда как в состоянии $|\mathbf{C}\rangle$ регистрации фотона не происходит. Отношение соответствующих вероятностей при этом равно $|1|^2 : |i|^2 = 1 : 1$; т. е. вероятности каждого из двух возможных событий равны, и фотон активирует фотоэлемент с той же вероятностью, с какой и вовсе не попадает на него.

Рассмотрим несколько более сложный случай. Допустим, что мы не заменяем зеркало в правом нижнем углу фотоэлементом, а полностью блокируем один из лучей неким непрозрачным «фотонопоглощающим» *препятствием* — скажем, луч, соответствующий состоянию $|\mathbf{D}\rangle$ фотона (см. рис. 5.13); при этом интерференция, имевшая место ранее, оказывается нарушена. Теперь, миновав последнее зеркало, фотон *может* перейти в состояние $|\mathbf{G}\rangle$ (возможность $|\mathbf{F}\rangle$ тоже пока никто не отменял) — однако лишь при условии, что *не* будет поглощен препятствием. Если препятствие *поглощает* фотон, то он вообще не дойдет до детекторов, ни в состоянии $|\mathbf{F}\rangle$, ни в состоянии $|\mathbf{G}\rangle$, ни в какой бы то ни было их комбинации. Если же поглощения не происходит, то последнего зеркала фотон достигнет, пребывая в «простом» состоянии $-|\mathbf{E}\rangle$, которое после прохождения зеркала эволюционирует в $-|\mathbf{F}\rangle - i|\mathbf{G}\rangle$. Таким образом, в конечном результате действительно присутствуют обе альтернативы — и $|\mathbf{F}\rangle$, и $|\mathbf{G}\rangle$.

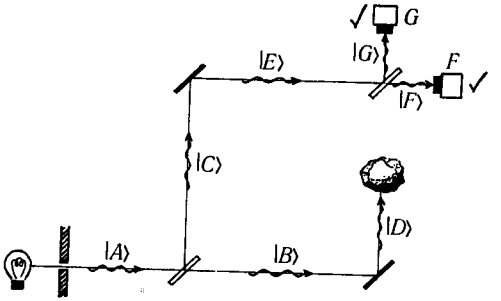



Рис. 5.13. Если перекрыть луч $|\mathbf{D}\rangle$ каким-либо препятствием, то детектор **G** также сможет зарегистрировать прибытие фотона (при условии, что этот фотон *не* будет раньше поглощен препятствием!).

В том случае, когда препятствие (в рассмотренной конкретной схеме) не поглощает фотон, комплексные весовые коэффициенты, соответствующие возможным состояниям $|\mathbf{F}\rangle$ и $|\mathbf{G}\rangle$, равны $-1$ и $-i$. Таким образом, отношение вероятностей равно $|-1|^2 : |-i|^2$, что опять дает одинаковые вероятности для обоих возможных событий — фотон активирует детектор в точке $|\mathbf{F}\rangle$ с той же вероятностью, с какой он активирует детектор в точке $|\mathbf{G}\rangle$.

Кроме того, само препятствие также следует считать «измерительным устройством» — коль скоро варианты «препятствие поглощает фотон» и «препятствие не поглощает фотон» мы рассматриваем как классические альтернативы, которым нельзя поставить в соответствие комплексные весовые коэффициенты. Даже если препятствие не устроено таким деликатным образом, что квантовое событие «поглощение препятствием фотона» порождает событие, наблюдаемое на классическом уровне, следует все же полагать, что такое устройство препятствия *принципиально возможно*. Существенным обстоятельством здесь является то, что в результате поглощения фотона некое значительное количество составляющего препятствие материала подвергается определенному, пусть и малому, возмущению — при этом практически невозможно собрать всю связанную с таким возмущением информацию, чтобы восстановить по ней сопутствующие



эффекты интерференции, характеризующие квантовые феномены. Итак, препятствие (во всяком случае, в практическом смысле) следует рассматривать как объект классического уровня, эквивалентный измерительному устройству — вне зависимости от того, регистрирует оно поглощение фотона каким-либо практически наблюдаемым образом или нет. (К этому вопросу мы еще вернемся, см. § 6.6.)

Учитывая вышесказанное, мы вольны воспользоваться «правилом квадратов модулей» и для вычисления вероятности того, что фотон и вправду окажется поглощен препятствием. Перед столкновением с препятствием фотон находится в состоянии $i|\mathbf{D}\rangle - |\mathbf{E}\rangle$, причем поглощается лишь фотон в состоянии $|\mathbf{D}\rangle$, тогда как в состоянии $|\mathbf{E}\rangle$ поглощения не происходит. Отношение вероятности поглощения к вероятности не-поглощения равно $|i|^2 : |-1|^2 = 1 : 1$ — обе альтернативы и здесь равновероятны.

Можно произвести еще одну небольшую модификацию рассматриваемой системы: уберем препятствие для луча **D**, зеркало же в правом нижнем углу не будем *заменять* детектором, но «прикрутим» вместо этого к зеркалу некое особого рода измерительное устройство. Предположим, что чувствительность этого устройства такова, что оно способно регистрировать (т. е. выводить на классический уровень) воздействие, оказываемое на зеркало фотоном при отражении, каким бы малым это воздействие ни было; сигналом о регистрации воздействия пусть будет отклонение стрелки на циферблате нашего устройства (см. рис. 5.14). Здесь отклонение стрелки вызывается фотоном в состоянии $|\mathbf{B}\rangle$, состояние же $|\mathbf{C}\rangle$ никакого воздействия на стрелку не оказывает. Принимая фотон в состоянии $|\mathbf{B}\rangle + i|\mathbf{C}\rangle$, устройство «коллапсирует волновую функцию» и интерпретирует суперпозицию *либо* как состояние $|\mathbf{B}\rangle$ (стрелка отклоняется), *либо* как состояние $|\mathbf{C}\rangle$ (стрелка остается неподвижной), причем вероятности обоих исходов одинаковы (поскольку $|1|^2 : |i|^2 = 1 : 1$). Таким образом, на *этом* этапе также имеет место процедура **R**. О дальнейшей судьбе фотона мы рассуждаем примерно так же, как мы делали это выше; при этом выясняется, что — как и в случае с препятствием — вероятности регистрации фотона детекторами **F** и **G** снова равны (причем независимо от того, отклонялась стрелка или нет). Для того чтобы фотон в данной схеме мог вызвать отклонение стрелки, зеркало в правом нижнем углу должно быть достаточно «подвижным», отсутствие же жесткого закрепления

нарушает хрупкий порядок, необходимый для возникновения той «деструктивной интерференции» между двумя траекториями движения фотонов от точки **A** к точке **G**, благодаря которой фотон в исходном примере не регистрировался детектором **G**.

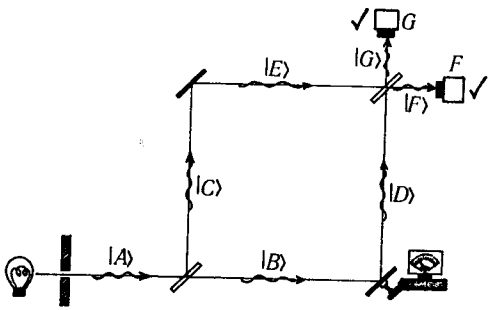



Рис. 5.14. Аналогичного эффекта можно достичь, поместив в правый нижний угол подвижное зеркало, снабженное неким детектором, который способен по движению зеркала определить, отразило оно фотон или нет. Интерференция здесь также оказывается нарушена, благодаря чему детектор в точке **G** получает возможность зарегистрировать прибытие фотона.

Читатель, должно быть, уже отметил некую досадную незавершенность всех наших рассуждений, выражающуюся в отсутствии ответа на вопрос «*Когда* (а главное, *почему*) квантовые правила переходят от квантового детерминизма комплексных весовых коэффициентов к классическим вероятностно-взвешенным недетерминированным альтернативам, каковой переход выражается математически в возведении в квадрат модулей соответствующих комплексных чисел?». Что есть такого в одних физических материальных образованиях — таких, например, как детекторы фотонов в точках **F** и **G** или зеркало в нижнем правом углу (или то же возможное препятствие для фотонов на пути луча **D**), — что делает их объектами классического уровня, в противоположность другим физическим объектам, скажем, фотонам, которые оказываются на квантовом уровне, и требуют поэтому совершенно иного с собой обращения? Только ли в том дело, что фотон —



это система физически простая, что позволяет рассматривать его целиком как объект квантового уровня, тогда как детекторы и препятствия являются системами сложными, которые можно рассматривать лишь приближенно, в результате чего тонкости квантового поведения растворяются в усредненных данных наблюдений? Многие физики, несомненно, ответят на последний вопрос утвердительно: *все* физические объекты, скажут они вам, следует рассматривать с позиций квантовой механики, и лишь руководствуясь соображениями удобства, мы исследуем большие и сложные системы классическими методами, причем правила вероятностей, задействованные в процедуре **R**, являются, в некотором роде, следствием упомянутого приближенного рассмотрения. В §§ 6.6 и 6.7 мы увидим, что от наших трудностей (связанных с присутствием в квантовой теории **X**-загадок) такая точка зрения отнюдь не спасает, равно как не объясняет она и смысла удивительного **R**-правила, согласно которому из квадратов модулей комплексных весовых коэффициентов чудесным образом получаются вероятности. И все же нам придется пока как-то усмирить нашу досаду и продолжить знакомство с выводами квантовой теории, в особенности с теми, что имеют отношение к ее **Z**-загадкам.

## 5.9. Решение задачи Элитцура – Вайдмана об испытании бомб

Мы уже знаем вполне достаточно для того, чтобы отыскать решение задачи об испытании бомб, поставленной в § 5.2. Прежде всего нужно выяснить, нельзя ли использовать сверхчувствительное зеркальце на носу бомбы в качестве измерительного устройства (как были использованы, например, препятствие и подвижное зеркало с детектором в описанных выше примерах). Построим систему зеркал (два непрозрачных, два полупрозрачных), которая в точности повторяет систему из предыдущего примера (см. рис. 5.14) за одним исключением: в правом нижнем углу вместо подвижного зеркала поместим зеркальце бомбы.

Смысл такого построения в том, что *если* бомба является холостой (в том единственном смысле, который подразумевается в условии задачи), то ее зеркальце остается в любом случае неподвижным (поскольку его заклинило), и общая картина эквивалентна показанной на рис. 5.12. Фотон, испущенный из источ-

ника, попадает на первое зеркало, будучи в состоянии $|\mathbf{A}\rangle$. Поскольку такая ситуация полностью совпадает с той, что мы рассмотрели в § 5.7, фотон после последнего зеркала приобретает, как и тогда, состояние $|\mathbf{F}\rangle$ (пропорциональное $|\mathbf{F}\rangle$, если точнее). Иначе говоря, детектор в точке **F** регистрирует прибытие фотона, а детектор в точке **G** не регистрирует ничего.

Если же бомба *исправна*, то падение фотона на ее зеркальце приводит к срабатыванию детонатора, и бомба взрывается. Бомба, фактически, представляет собой измерительное устройство. Альтернативы квантового уровня — «фотон падает на зеркальце» и «фотон не падает на зеркальце» — переводятся бомбой в альтернативы классического уровня — «бомба взрывается» и «бомба не взрывается». На состояние $|\mathbf{B}\rangle + i|\mathbf{C}\rangle$ бомба реагирует взрывом, если обнаруживает, что фотон находится в состоянии $|\mathbf{B}\rangle$; если же фотон находится в каком-то ином состоянии (т. е., в данном случае, $|\mathbf{C}\rangle$), бомба не взрывается. Отношение вероятностей этих двух событий равно $|1|^2 : |i|^2 = 1 : 1$. Если бомба таки взорвалась, это означает, что она зарегистрировала прибытие фотона, а что будет дальше, никого уже не интересует. Если же взорваться бомбе не удалось, то состояние фотона редуцируется (как результат процедуры **R**) до состояния $i|\mathbf{C}\rangle$ (падение на зеркало в левом верхнем углу), сменяясь далее (после отражения от этого зеркала) состоянием $-|\mathbf{E}\rangle$. По прохождении последнего (полупрозрачного) зеркала фотон переходит в состояние $-|\mathbf{F}\rangle - i|\mathbf{G}\rangle$, т. е. отношение вероятностей возможных исходов — «прибытие фотона регистрируется детектором в точке **F**» и «прибытие фотона регистрируется детектором в точке **G**» — равно $|-1|^2 : |-i|^2 = 1 : 1$. Точно такое же отношение мы получили в примерах, описанных в предыдущем параграфе, для тех случаев, когда фотон не поглощался препятствием, а стрелка не отклонялась. Детектор, расположенный в точке **G**, получает, таким образом, вполне определенную возможность уловить фотон.

Предположим теперь, что при проведении одного из таких испытаний в некоторых случаях «не-взрыва» бомбы обнаруживается, что детектор **G** и в самом деле регистрирует прибытие фотона. Согласно нашим рассуждениям, это возможно лишь в том случае, если детонатор бомбы *исправен*! Если бомба неисправна, то фотон может быть зарегистрирован только детектором **F**. Следовательно, во всех случаях, когда срабатывает детектор **G**, мы можем с чистой совестью гарантировать, что данная бомба



«работоспособна» и в случае необходимости не подведет. Таким образом, задачу об испытании бомб (§ 5.2) можно считать решенной[6].

Судя по участвующим в процессе вероятностям, после достаточно большого количества испытаний половина бомб взорвется, и никакой дальнейшей пользы из них извлечь не удастся. Более того, на тех бомбах, что не взорвались, детектор **G** сработает только в половине случаев. Таким образом, после того, как мы переберем все бомбы одну за другой, мы сможем *гарантировать* работоспособность только четверти из первоначального запаса исправных бомб. Оставшиеся бомбы мы можем подвергнуть повторному испытанию, отбирая те, на которых сработал детектор **G**. Повторим испытание еще раз. И еще. В конечном счете у нас останется треть (поскольку $\frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \ldots = \frac{1}{3}$) от первоначального количества исправных бомб, но зато *все* эти бомбы будут гарантированно работоспособны. (Я не знаю, для чего эти бомбы предназначены, однако, думаю, благоразумно будет лишних вопросов не задавать!)

[6] ***Shabbos-ключ, или Субботний выключатель.*** Тот факт, что и Элитцур, и Вайдман работают в университетах Израиля, натолкнул нас с Артуром Экертом однажды во время беседы на идею создания устройства для помощи тем евреям, кто строго соблюдает все установления иудаизма и кому, следовательно, запрещается включать или выключать электрические приборы в субботу. Мы могли бы запатентовать соответствующее устройство и заработать тем самым целое состояние, однако вместо этого решили сделать нашу эпохальную идею достоянием общественности, дабы ею мог воспользоваться любой еврей, у которого возникнет в таком устройстве потребность. Для создания устройства понадобится источник, способный испускать непрерывную последовательность фотонов, два полупрозрачных и два непрозрачных зеркала и фотоэлемент, соединенный с прибором, который необходимо включать/выключать. Схема аналогична изображенной на рис. 5.13, фотоэлемент помещается в точке **G**. Для того чтобы включить или выключить прибор, следует поместить палец на пути луча **D**, приблизительно там же, где на рис. 5.13 находится препятствие. Если фотон падает на палец, то ничего не происходит — разумеется, никакого греха в этом нет. (Фотоны и без того постоянно бомбардируют наши пальцы, и по субботам с ничуть не меньшим усердием.) Если же палец с фотоном не встретится, то имеется 50%-я вероятность (буде на то воля Божия), что обслуживаемый устройством электроприбор включится. Несомненно, не будет греха и в том, что фотон упадет *не* на ваш палец, а на выключатель прибора. (Тут имеется, правда, одно возражение практического свойства: источники, способные испускать по одному фотону, весьма сложны — и дороги. Однако особой необходимости в них, в сущности, нет. Сгодится любой источник фотонов, поскольку приведенное выше рассуждение применимо и к каждому отдельному фотону из пучка.)

Читателю описанная процедура может показаться чересчур расточительной, однако поразительно здесь то, что она вообще осуществима. Никакими классическими методами задача не решается. Только в квантовой теории контрфактуальные вероятности могут действительно повлиять на физический результат. Наша квантовая процедура позволяет добиться того, что кажется невозможным, — что и в самом деле невозможно в рамках классической физики. Следует, кроме того, отметить, что с помощью некоторых усовершенствований потери можно снизить с двух третей до практически половины (см. [114]). Еще более поразительного результата добились не так давно П. Г. Квят, Х. Вайнфуртер, А. Цайлингер и М. Казевич, описав процедуру (отличную от решения Элитцура – Вайдмана), позволяющую снизить потери почти до нуля!

Что касается сложностей с разработкой экспериментального устройства, способного испускать отдельные фотоны по одному за раз, то они теперь позади — такие устройства уже созданы и вполне доступны (см. [168]).

В заключение отмечу, что в качестве измерительного устройства вовсе не обязательно должен выступать столь «сногсшибательный» объект, как фигурирующая в условии задачи бомба. Более того, нет никакой необходимости в том, чтобы упомянутое «устройство» оповещало бы весь внешний мир о том, что оно зарегистрировало (или не зарегистрировало) прибытие фотона. Подвижное зеркало может само по себе послужить измерительным устройством, если его вес достаточно мал для того, чтобы оно могло сколько-нибудь заметно поворачиваться под воздействием падающих на него фотонов и затем останавливаться вследствие трения. Один лишь факт подвижности зеркала (скажем, зеркала в правом нижнем углу, как в рассмотренном примере) позволит детектору в точке **G** зарегистрировать прибытие фотона, даже если зеркало в действительности и *не* повернулось, указывая тем самым на то, что фотон отправился другой дорогой. Достичь точки **G** фотону позволяет *потенциальная возможность* поворота зеркала и ничто иное! Очень похожую роль играет и поглощающее фотоны препятствие из предыдущего параграфа. Оно, в сущности, служит для «измерения» наличия фотона где-то на пути, описываемом последовательными состояниями $|\mathbf{B}\rangle$ и $|\mathbf{D}\rangle$. То, что препятствие не поглощает фотон, будучи на это способно,



является точно таким же «измерением», каким мы считаем состоявшееся поглощение фотона.

Такие отрицательные и бесконтактные измерения, называемые *нулевыми* (или невзаимодействующими) измерениями (см. [91]), имеют большое теоретическое (а возможно, в конечном счете, и практическое) значение. Предсказания квантовой теории относительно такого рода ситуаций непосредственно подтверждаются экспериментально. В частности, Квят, Вайнфуртер и Цайлингер разработали и провели эксперимент, *точно* воспроизводящий теоретическую процедуру Элитцура – Вайдмана для решения задачи об испытании бомб! И теоретические ожидания полностью подтвердились, что, впрочем, нас уже почему-то не удивляет. Сами же нулевые измерения мы по праву относим к наиболее фундаментальным **Z**-загадкам квантовой теории.

## 5.10. Квантовая теория спина. Сфера Римана

Для того, чтобы разобраться со второй вводной квантовой головоломкой, необходимо рассмотреть структуру квантовой теории несколько подробнее. Если помните, в центр моего додекаэдра (равно как и додекаэдра моего коллеги) был помещен атом со спином $\frac{3}{2}$. Что же такое спин, и каково его место в квантовой теории?

Спин — неотъемлемое свойство частицы. По существу, физическое понятие спина совпадает с понятием вращения[7] (или *кинетического момента*) классического объекта — например, бильярдного шара, футбольного мяча или даже планеты Земля. Существует, впрочем, различие (незначительное): наибольший (практически весь) вклад в кинетический момент макроскопического объекта дают круговые движения всех составляющих его частиц вокруг общего центра масс, тогда как спин одной-единственной частицы есть свойство, присущее самой частице. Более того, спин элементарной частицы обладает любопытной особенностью: его *величина* всегда *одинакова*, а вот направление оси спина может быть разным (хотя, надо сказать, что эта самая «ось» также ведет себя весьма странно, в общем случае малосообразно с тем, как ведут себя классические оси враще-

[7] Английское *spin* как раз и означает, среди прочего, «вращение». — *Прим. перев.*

ния). Спин измеряется в единицах фундаментальной квантовомеханической постоянной $\hbar$; символ этот предложен Дираком для обозначения величины, равной постоянной Планка $h$, деленной на $2\pi$. Спин частицы всегда равен (неотрицательному) целому или полуцелому кратному постоянной $\hbar$: $0$, $\frac{1}{2}\hbar$, $\hbar$, $\frac{3}{2}\hbar$, $2\hbar$ и т. д. Мы, соответственно, говорим: частица со спином $0$, $\frac{1}{2}$, $1$, $\frac{3}{2}$, $2$ и т. д.

Начнем с рассмотрения простого случая: спин $\frac{1}{2}$; таким спином обладают, например, электрон и нуклоны (протон и нейтрон). (Спин 0 мы рассматривать не будем, поскольку он *слишком* прост — в этом случае спин может находиться лишь в одном, сферически симметричном, состоянии.) Все состояния спина $\frac{1}{2}$ являются линейными суперпозициями двух состояний: скажем, правого спина вокруг оси, направленной вертикально *вверх* (обозначим это состояние через $|\uparrow\rangle$) и правого спина вокруг оси, направленной вертикально *вниз* (обозначим $|\downarrow\rangle$); см. рис. 5.15. Таким образом, в общем случае состояние спина можно представить в виде комплексной комбинации $|\psi\rangle = w|\uparrow\rangle + z|\downarrow\rangle$. На практике же каждой такой комбинации соответствует вполне определенное состояние спина (величины $\frac{1}{2}\hbar$) частицы, при котором отношение комплексных коэффициентов $w$ и $z$ определяет направление оси спина. Выбор направлений $\uparrow$ и $\downarrow$ достаточно условен: для однозначного описания состояния спина сгодилась бы и любая другая пара направлений.

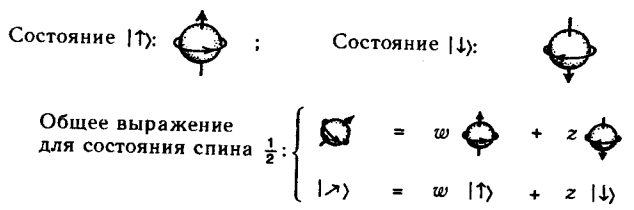



Рис. 5.15. В случае частицы со спином $\frac{1}{2}$ (электрона, протона или нейтрона) все спиновые состояния представляют собой комплексные суперпозиции двух основных состояний: «вверх» и «вниз».



Попробуем представить все вышесказанное в более явном и геометрически наглядном виде. Такое представление поможет нам увидеть, что комплексные весовые коэффициенты $w$ и $z$ вовсе не являются такими уж абстрактными конструкциями, какими они могли показаться на первый взгляд. Более того, к геометрии пространства они имеют самое непосредственное отношение. (Мне думается, такие геометрические воплощения понравились бы Кардано и, возможно, облегчили бы его «мучения разума» — впрочем, и квантовая теория вполне исправно снабжает наши разумы все новыми мучениями!)

Для начала будет весьма полезно ознакомиться со ставшим уже стандартным представлением комплексных чисел в виде точек на плоскости. (У этой плоскости много названий: плоскость Арганда, плоскость Гаусса, плоскость Весселя или просто *комплексная* плоскость.) Идея состоит в том, чтобы поставить в соответствие комплексному числу $z = x + iy$ (где $x$ и $y$ — вещественные числа) точку, координаты которой в некоторой заданной прямоугольной системе координат равны $(x, y)$ (см. рис. 5.16). Таким образом, например, четыре комплексных числа $1$, $1 + i$, $i$ и $0$ образуют на комплексной плоскости квадрат. Существуют простые геометрические правила для отыскания суммы и произведения двух комплексных чисел (см. рис. 5.17). Отрицательное комплексное число $-z$ находится отражением точки, соответствующей числу $z$, относительно начала координат; комплексное сопряженное $\bar{z}$ — отражением точки $z$ относительно оси $x$.

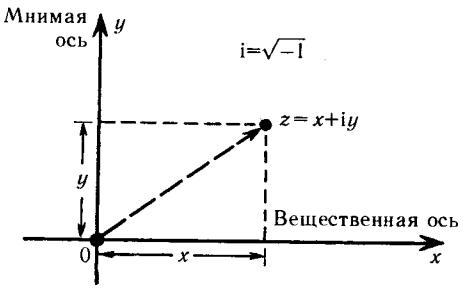



Рис. 5.16. Представление комплексного числа в виде точки на комплексной плоскости (плоскости Арганда — Гаусса — Весселя).

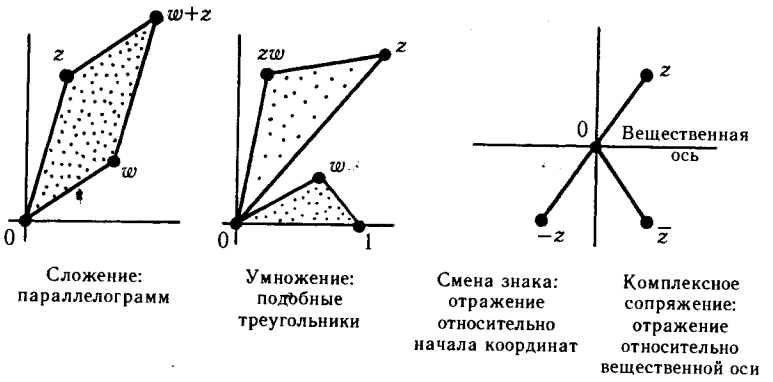



Рис. 5.17. Геометрические описания основных операций над комплексными числами.

Модуль комплексного числа равен расстоянию от соответствующей этому числу точки до начала координат; квадрат модуля, таким образом, равен квадрату этого расстояния. Точки, расстояние от которых до начала координат равно единице, образуют *единичную окружность* (см. рис. 5.18). Этим точкам соответствуют комплексные числа с *единичным модулем*, называемые иногда *чистыми фазами*; эти числа можно записать в виде

$$e^{i\theta} = \cos\theta + i\sin\theta,$$

здесь $\theta$ — вещественное число, равное величине угла между прямой, соединяющей начало координат с соответствующей этому числу точкой, и осью $x$.[8]

Теперь выясним, как в таком представлении выглядят *отношения* комплексных чисел. Выше я уже указывал на то, что при умножении вектора состояния на ненулевое комплексное число состояние не претерпевает физических изменений (например, если помните, состояния $-2|\mathbf{F}\rangle$ и $|\mathbf{F}\rangle$ мы полагали физически одинаковыми). Таким образом, в общем случае, состояние $|\psi\rangle$ физически идентично состоянию $u|\psi\rangle$ при любом ненулевом комплексном $u$. Применительно к состоянию

$$|\psi\rangle = w|\uparrow\rangle + z|\downarrow\rangle,$$

умножение $w$ и $z$ на одно и то же ненулевое комплексное число $u$ не приведет к какому-либо изменению физического феномена, соответствующего этому состоянию. Физически различными спиновые состояния могут быть только в том случае, если их векторы состояний характеризуются *различными отношениями* $z : w$ (а при $u \neq 0$ отношения $uz : uw$ и $z : w$ равны).

[8] Вещественное число $e$ называется «основанием натурального логарифма»: $e = 2,7182818285\ldots$. Запись $e^z$ означает «число $e$ в степени $z$»; для вычисления значения такого выражения используют следующее разложение:

$$e^z = 1 + z + \frac{z^2}{1\times 2} + \frac{z^3}{1\times 2\times 3} + \frac{z^4}{1\times 2\times 3\times 4} + \ldots.$$



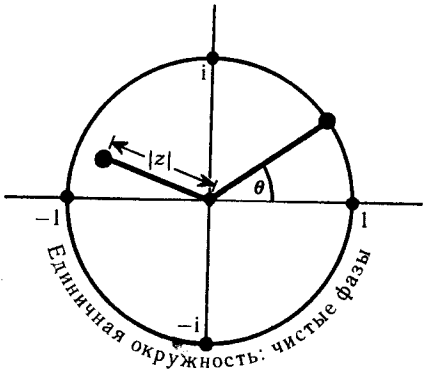



Рис. 5.18. Единичную окружность образуют точки, соответствующие комплексным числам $z = e^{i\theta}$, где $\theta$ — вещественное число; $|z| = 1$.

Как же изобразить комплексное отношение геометрически? Существенное отличие комплексного отношения от просто комплексного числа заключается в том, что в качестве значения комплексного отношения допускается не только конечное комплексное число, но и *бесконечность* (обозначается символом $\infty$). Так, если рассматривать, в общем случае, отношение $z : w$ как эквивалент «одиночного» комплексного числа $z/w$, то при $w = 0$ мы сталкиваемся с некоторыми, мягко говоря, затруднениями. Для того чтобы этих затруднений избежать, математики условились в случае $w = 0$ полагать число $z/w$ равным бесконечности. Такая

ситуация возникает, например, в состоянии «спин вниз»: $|\psi\rangle = z|\downarrow\rangle = 0|\uparrow\rangle + z|\downarrow\rangle$. Вспомним, что нулю не могут быть равны *оба* коэффициента (т. е. и $w$, и $z$ одновременно), поэтому случай $w = 0$ вполне допустим. (Мы могли бы вместо $z/w$ взять отношение $w/z$, если оно по каким-либо причинам понравилось бы нам больше; тогда символ $\infty$ понадобился бы нам для случая $z \neq 0$, что соответствует состоянию «спин вверх». Никакой разницы между этими двумя описаниями нет.)

Пространство всех возможных комплексных отношений мы можем представить с помощью так называемой *сферы Римана*. Точки, образующие сферу Римана, соответствуют комплексным числам, либо $\infty$. Сферу Римана можно изобразить в виде единичной сферы, экваториальная плоскость которой совпадает с комплексной плоскостью, а центр располагается в точке начала координат (т. е. в нуле). Собственно экватор сферы есть не что иное, как единичная окружность на комплексной плоскости (см. рис. 5.19). Для представления какого-либо комплексного отношения, скажем, $z : w$, мы отмечаем на комплексной плоскости точку $P$, соответствующую комплексному числу $p = z/w$ (допустим пока, что $w \neq 0$), а затем проецируем эту точку $P$ в точку $P'$ на сфере, при этом в качестве центра проекции выбираем *южный полюс* $S$ сферы. Иначе говоря, мы проводим через точки $S$ и $P$ прямую; там, где эта прямая пересекает сферу (кроме самой точки $S$), отмечаем точку $P'$. Такое точечное отображение плоскости на сферу называется *стереографической проекцией*. Сам южный полюс $S$ при таком отображении соответствует комплексному отношению $\infty$. В самом деле, представим себе, что точка $P$ комплексной плоскости удалена на очень большое расстояние от центра координат; соответствующая ей точка $P'$ на сфере окажется при этом очень близко от полюса $S$ — в пределе, когда модуль комплексного числа $p$ устремляется к бесконечности, точки $P'$ и $S$ совпадают.

Сфера Римана играет фундаментальную роль в квантовом описании систем с двумя состояниями. Эта роль не всегда очевидна, однако это не делает ее менее важной, и сфера Римана, пусть и незримо, где-то на сцене все равно присутствует. Она описывает — в абстрактном геометрическом виде — пространство всех физически достижимых состояний, которые можно получить из двух различных квантовых состояний посредством квантовой линейной суперпозиции. В качестве исходных



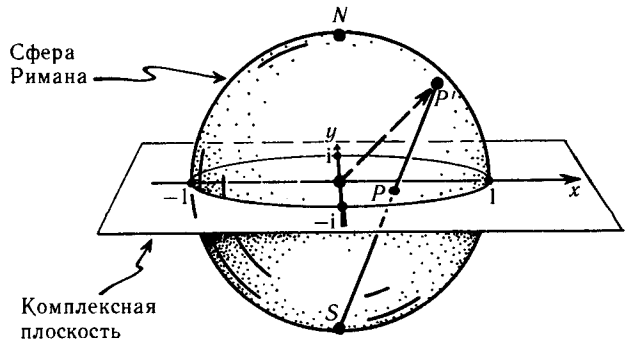



Рис. 5.19. Сфера Римана. Точка $P$ на комплексной плоскости, соответствующая числу $p = z/w$, проецируется из южного полюса $S$ на точку $P'$ на сфере. Направление $\overrightarrow{OP}$ совпадает с направлением оси спина для общего состояния спина $\frac{1}{2}$ (см. рис. 5.15).

можно взять, например, возможные состояния фотона $|\mathbf{B}\rangle$ и $|\mathbf{C}\rangle$. В общем случае их линейная комбинация имеет вид $w|\mathbf{B}\rangle + z|\mathbf{C}\rangle$. В § 5.7 мы подробно рассматривали только один конкретный случай $|\mathbf{B}\rangle + i|\mathbf{C}\rangle$ (результат отражения/пропускания света, падающего на полусеребрёное зеркало), однако нетрудно реализовать и другие комбинации состояний. Для этого нужно всего лишь изменить степень «серебрёности» зеркала и поместить на пути одного из лучей что-нибудь преломляющее. Так можно набрать полную сферу Римана всевозможных альтернативных состояний, соответствующих различным физическим ситуациям вида $w|\mathbf{B}\rangle + z|\mathbf{C}\rangle$, т. е. комбинациям двух начальных состояний $|\mathbf{B}\rangle$ и $|\mathbf{C}\rangle$.

Впрочем, в таких случаях геометрическая роль сферы Римана как раз и неочевидна. Однако возможны и иные ситуации, в которых целесообразность построения сферы Римана проявляется в полной мере. Самым наглядным примером такого рода является описание спиновых состояний частицы со спином $\frac{1}{2}$ — электрона, скажем, или протона. В общем случае спиновое состояние можно записать в виде комбинации

$$|\psi\rangle = w|\uparrow\rangle + z|\downarrow\rangle;$$

как оказывается (при соответствующем выборе направлений $\uparrow$ и $\downarrow$ из физически эквивалентных возможных вариантов), это самое $|\psi\rangle$ представляет собой состояние правого спина (величины $\frac{1}{2}\hbar$), направление оси которого совпадает с направлением от начала координат к точке, соответствующей отношению $z/w$, на сфере Римана. Таким образом, любое направление в пространстве выступает как возможное направление оси спина для любой частицы со спином $\frac{1}{2}$. Хотя бóльшая часть спиновых состояний представляется изначально в виде «таинственных комплексно-взвешенных комбинаций возможных альтернативных состояний» (т. е. состояний $|\uparrow\rangle$ и $|\downarrow\rangle$), мы видим, что эти состояния ничуть не более (но и не менее) таинственны, чем оригинальные состояния $|\uparrow\rangle$ и $|\downarrow\rangle$, выбранные нами в качестве начальных. Каждое физически реально в той же мере, что и все остальные.

А что же с состояниями большего спина? Здесь ситуация становится несколько более запутанной — и *более* таинственной! Приводимое ниже общее описание не пользуется широкой известностью среди современных физиков, хотя оно было предложено еще в 1932 году блестящим итальянским физиком Этторе Майораной (в 1938 году, в возрасте 31 года, Майорана бесследно исчез с борта входившего в Неаполитанский залив парома при обстоятельствах, которые до сих пор не получили удовлетворительного объяснения).

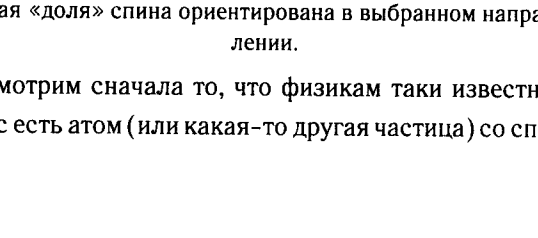



Рис. 5.20. Измерение спина с помощью установки Штерна — Герлаха. Для частицы со спином $\frac{1}{2}n$ мы можем получить $n+1$ возможных результатов, в зависимости от того, какая «доля» спина ориентирована в выбранном направлении.

Рассмотрим сначала то, что физикам таки известно. Допустим, у нас есть атом (или какая-то другая частица) со спином $\frac{1}{2}n$.



В качестве исходного направления мы снова можем выбрать направление вверх, а заодно и полюбопытствуем, «какая доля» спина атома действительно ориентирована в этом направлении (т. е. является правой относительно направленной вверх оси). Для удовлетворения любопытства можно воспользоваться стандартным устройством, которое называется установкой Штерна — Герлаха и способно осуществлять упомянутые измерения с помощью неоднородного магнитного поля. Как выясняется, различных возможных вариантов развития событий всего $n+1$, что обусловлено тем фактом, что атомы в магнитном поле могут отклоняться только в одном из $n+1$ возможных направлений (см. рис. 5.20). Доля спина, ориентированного в выбранном направлении, определяется конкретным направлением, в котором отклоняется атом. Будучи измеренной в единицах $\frac{1}{2}\hbar$, доля ориентированного в данном направлении спина принимает одно из следующих значений: $n$, $n-2$, $n-4$, ..., $2-n$, $-n$. Возможные же спиновые состояния для атома со спином $\frac{1}{2}n$ представляют собой комплексные суперпозиции перечисленных допустимых состояний. Возможные результаты измерения Штерна — Герлаха для спина $n+1$ (направление поля в установке — вертикально вверх) я буду записывать следующим образом:

$$|\uparrow\uparrow\uparrow\ldots\uparrow\rangle,\quad |\downarrow\uparrow\uparrow\ldots\uparrow\rangle,\quad |\downarrow\downarrow\uparrow\ldots\uparrow\rangle,\quad \ldots,\quad |\downarrow\downarrow\downarrow\ldots\downarrow\rangle,$$

что соответствует значениям $n$, $n-2$, $n-4$, ..., $2-n$, $-n$ доли спина, ориентированного в этом направлении (запись каждого состояния содержит ровно $n$ стрелок). Результаты можно интерпретировать так: каждая стрелка вверх дает долю $\frac{1}{2}\hbar$ спина, ориентированного вверх, а каждая стрелка вниз дает долю $\frac{1}{2}\hbar$ спина, ориентированного вниз. Складывая эти величины, мы получаем полный спин для каждого конкретного случая измерения с помощью установки Штерна — Герлаха (при ориентации осей в направлении вверх/вниз).

В общем случае суперпозиция этих состояний записывается в виде комплексной комбинации

$$z_0|\uparrow\uparrow\uparrow\ldots\uparrow\rangle + z_1|\downarrow\uparrow\uparrow\ldots\uparrow\rangle + z_2|\downarrow\downarrow\uparrow\ldots\uparrow\rangle + \ldots + z_n|\downarrow\downarrow\downarrow\ldots\downarrow\rangle,$$

где хотя бы один из комплексных коэффициентов $z_0, z_1, z_2, \ldots, z_n$

не равен нулю. Можно ли представить такое состояние с помощью отдельных направлений оси спина, отличных от элементарных «вверх» или «вниз»? Как показал Майорана, такое представление действительно возможно, однако следует допустить, что направления эти будут вполне независимы друг от друга: нет никакой необходимости брать в качестве исходных обязательно пару обязательно противоположных направлений (как в случае измерения с помощью установки Штерна – Герлаха). Иными словами, общее состояние спина $\frac{1}{2}n$ мы представим в виде набора из $n$ независимых «стрелок-направлений»; эти направления можно рассматривать как направления, задаваемые $n$ точками на сфере Римана, — при этом каждая «стрелка» исходит из начала координат и заканчивается в соответствующей точке на сфере (см. рис. 5.21). Важно помнить, что мы имеем дело с *неупорядоченной* совокупностью точек (или направлений), и, следовательно, в порядок их рассмотрения никакого особого смысла вкладывать не нужно.

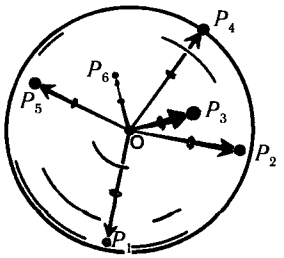



Рис. 5.21. Майорана описывает общее состояние спина $\frac{1}{2}n$ как неупорядоченную совокупность из $n$ точек $P_1, P_2, \ldots, P_n$ на сфере Римана, причем каждая точка соответствует «элементарному» спину $\frac{1}{2}$, направление оси которого совпадает с направлением от начала координат к этой самой точке.

Получившаяся картина выглядит очень странно — если мы попытаемся подойти к квантовомеханическому спину с теми же мерками, что и к привычной концепции вращения на классическом уровне. Вращение классического объекта (например, бильярдного шара) всегда происходит вокруг некоторой вполне определенной оси, тогда как объекту квантового уровня позволено, судя по всему, вращаться одновременно вокруг множества осей, ориентированных в самых разных направлениях. Полагая, что квантовые объекты — это, в сущности, те же классические объекты, только «маленькие», мы неизбежно сталкиваемся с парадоксом. Чем больше величина спина, тем большее количество направлений осей необходимо для описания его состояния. Почему же, в таком случае, классические объекты не вращаются вокруг нескольких осей одновременно? Перед нами типичный пример квантовой **X**-загадки. Что-то вмешивается в процесс (на некоем неустановленном уровне), и мы обнаруживаем, что большинство типов квантовых состояний на классическом уровне феноменов — т. е. там, где мы могли бы их воспринимать, — не возникают вовсе (или, по меньше мере, почти никогда). В случае спина мы видим, что на классическом уровне сохраняются только те состояния, в которых оси преимущественно группируются в каком-то одном направлении — в направлении оси вращения классического вращающегося объекта.



В квантовой теории есть одно занимательное допущение, называемое «принципом соответствия». Суть этого принципа такова: как только какая-либо физическая величина (например, величина спина) возрастает до некоего предела, становится *возможным* такое поведение системы, которое очень близко аппроксимирует классическое поведение (как, например, спиновое состояние, где направления всех осей приблизительно одинаковы). Однако нигде почему-то не объясняется, каким образом к подобным состояниям приводит одна лишь шрёдингерова эволюция **U**. В действительности «классические состояния» *так* не возникают почти никогда. Состояния классического типа являются результатом действия совершенно иной процедуры — редукции **R** вектора состояния.

## 5.11. Местонахождение частицы и ее количество движения

Еще более наглядным примером такого рода является квантовомеханическая концепция *положения* частицы в пространстве. Выше мы говорили о том, что состояние частицы может

включать в себя суперпозицию двух или более различных ее положений. (Вспомним также и о примерах из § 5.7, где после прохождения полупрозрачного зеркала фотон оказывается в состоянии, предполагающем его нахождение в двух различных лучах одновременно.) Такие суперпозиции возможны и в случае любых других типов частиц (как простых, так и составных) — электронов, протонов, атомов или молекул. Более того, в части **U** формализма квантовой теории нет ничего, что запрещало бы оказаться в двусмысленном состоянии суперпозиции положений макроскопическим объектам вроде бильярдных шаров. Однако никто ни разу не видел бильярдный шар в состоянии суперпозиции нескольких положений одновременно, равно как никто не видел и бильярдный шар, вращающийся одновременно вокруг нескольких осей. Почему получается так, что некоторые физические объекты оказываются слишком большими, или слишком массивными, или слишком какими-то еще для того, чтобы «протиснуться» на квантовый уровень, вследствие чего не могут в реальном мире находиться в какой бы то ни было суперпозиции состояний? В стандартной квантовой теории переход от квантовых суперпозиций возможных альтернатив к единственному действительному классическому результату осуществляется исключительно благодаря действию процедуры **R**. Действие же одной лишь процедуры **U** практически неизбежно приводит к таким классическим суперпозициям, которые выглядят, мягко говоря, «неестественно». (К этому вопросу я еще вернусь в § 6.1.)

На квантовом же уровне те состояния частицы, в которых она не имеет четко определенного положения, могут играть, ни много ни мало, фундаментальную роль: если частица обладает определенным *количеством движения* (т. е. движется по некоторой определенной траектории в определенном направлении, а не в суперпозиции нескольких разных направлений одновременно), то в состоянии этой частицы непременно должна присутствовать суперпозиция всех ее различных *положений* одновременно. (Это одно из свойств уравнения Шрёдингера, и для должного объяснения этого свойства потребовалось бы слишком далеко углубиться в технические детали, что нам сейчас совсем не нужно; см., например, НРК, с. 243–250, а также [94] и [70]. Оно, кроме того, тесно связано с *принципом неопределенности* Гейзенберга, устанавливающим предел точности для одновременного измерения положения частицы и ее количества движения.) Более того,



в состояниях с определенным количеством движения частицы демонстрируют колебательное (в направлении движения) пространственное поведение, чего при обсуждении состояний фотонов в § 5.7 мы не учитывали. Строго говоря, термин «колебательное» здесь не совсем подходит. Как выясняется, упомянутые «колебания» отнюдь не похожи на колебания, скажем, струны — комплексные весовые коэффициенты не «мечутся» взад и вперед сквозь начало координат на комплексной плоскости, но, будучи чистыми фазами (см. рис. 5.18), движутся вокруг начала координат с постоянной скоростью, причем эта самая скорость задает частоту $\nu$, пропорциональную энергии $E$ частицы в соответствии со знаменитой формулой Планка $E = h\nu$. (Графическое представление состояний количества движения в виде этакого «штопора» можно найти в НРК, рис. 6.11.) Все эти вещи, хоть они и важны для квантовой теории, в наших дальнейших рассуждениях особой роли не играют, поэтому читатель вполне может обойтись и без детального их изучения.

В общем случае комплексные весовые коэффициенты вовсе не обязательно должны иметь именно такой «колебательный» вид, они могут изменяться от точки к точке произвольным образом. Весовые коэффициенты задают комплексную функцию положения, которая называется *волновой функцией* частицы.

## 5.12. Гильбертово пространство

Чтобы более внятно (и более точно) рассказать о том, как работает процедура **R** в стандартных квантовомеханических описаниях, необходимо перейти на несколько (совсем немного) более высокий уровень математической абстракции. Семейство всех возможных состояний квантовой системы образует так называемое *гильбертово пространство*. Нужды объяснять значение этого термина во всех математических тонкостях у нас в данный момент нет, однако некоторое представление о нем все же получить стоит — это поможет нам прояснить существующую картину квантового мира.

Первая и наиболее важная особенность, на которую следует обратить внимание: гильбертово пространство является *комплексным векторным пространством*. Это, в сущности, означает, что здесь мы вправе выполнять действия с комплексно-взвешенными комбинациями, посредством которых описываются

квантовые состояния. Для обозначения элементов гильбертова пространства я продолжу использовать диракову скобку «кет», т. е. если состояния $|\psi\rangle$ и $|\phi\rangle$ являются элементами гильбертова пространства, то таким же его элементом является и состояние $w|\psi\rangle + z|\phi\rangle$, где $w$ и $z$ — любая пара комплексных чисел. Допускается даже комбинация $w = z = 0$, она дает элемент $\mathbf{0}$ гильбертова пространства — единственный элемент, не соответствующий никакому возможному физическому состоянию. Как и в любом другом векторном пространстве здесь действуют самые обыкновенные алгебраические правила:

$$\begin{aligned}
|\psi\rangle + |\phi\rangle &= |\phi\rangle + |\psi\rangle,\\
|\psi\rangle + (|\phi\rangle + |\chi\rangle) &= (|\psi\rangle + |\phi\rangle) + |\chi\rangle,\\
w(z|\psi\rangle) &= (wz)|\psi\rangle,\\
(w+z)|\psi\rangle &= w|\psi\rangle + z|\psi\rangle,\\
z(|\psi\rangle + |\phi\rangle) &= z|\psi\rangle + z|\phi\rangle,\\
0|\psi\rangle &= \mathbf{0},\\
z\mathbf{0} &= \mathbf{0},
\end{aligned}$$

а это более или менее означает, что мы можем использовать алгебраическую систему обозначений привычным нам образом.

Иногда гильбертово пространство имеет конечную размерность — как, например, при описании спиновых состояний частицы. В случае спина $\frac{1}{2}$ гильбертово пространство двумерно, а его элементы представляют собой комплексные линейные комбинации двух состояний, $|\uparrow\rangle$ и $|\downarrow\rangle$. Для спина $\frac{1}{2}n$ гильбертово пространство $(n+1)$-мерно. Однако размерность гильбертова пространства может быть и *бесконечной* — такое пространство необходимо, например, для описания состояний положения частицы. В этом случае каждое альтернативное положение, которое может занимать частица, рассматривается как отдельное измерение гильбертова пространства. Общее же состояние, определяющее квантовое местоположение частицы, записывается как комплексная суперпозиция *всех* этих различных отдельных положений (волновая функция для данной конкретной частицы). Надо сказать, что с рассмотрением такого бесконечномерного гильбертова пространства связаны определенные математические осложнения, которые лишь запутают нас без всякой на то необходимости, поэтому ниже я сосредоточусь (в основном) на конечномерном случае.



Попытавшись представить гильбертово пространство визуально, мы сталкиваемся с двумя трудностями. Во-первых, размерность такого пространства, как правило, слишком велика для того, чтобы наше воображение сколько-нибудь адекватно справилось с задачей. Во-вторых, пространство это является не вещественным, но *комплексным*. Впрочем, часто бывает полезно не задумываться о подобных трудностях с самого начала — это помогает выработать некоторое интуитивное понимание математических аспектов концепции. Поэтому давайте на некоторое время сделаем вид, будто для представления гильбертова пространства вполне достаточно той привычной двух- или трехмерной картины, которая у нас уже есть. На рис. 5.22 проиллюстрирована геометрически операция линейной суперпозиции на примере обычного трехмерного пространства.

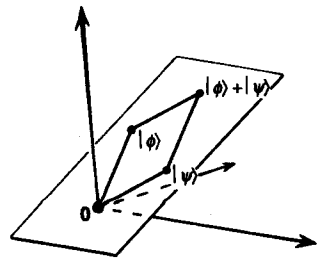



Рис. 5.22. Если вообразить, что гильбертово пространство тождественно трехмерному евклидову пространству, то сумму векторов $|\psi\rangle$ и $|\phi\rangle$ можно найти с помощью обычного правила параллелограмма (в плоскости $(\mathbf{0}, |\psi\rangle, |\phi\rangle)$).

Вспомним, что вектор квантового состояния $|\psi\rangle$ соответствует тому же физическому состоянию, что и любой кратный ему вектор $u|\psi\rangle$, где $u$ — ненулевое комплексное число. В нашей геометрической интерпретации это означает, что физическое состояние представляется не одинокой точкой в гильбертовом пространстве, но прямой, соединяющей гильбертову точку $|\psi\rangle$ с началом координат $\mathbf{0}$ (такую прямую называют *лучом*). При-

мер луча изображен на рис. 5.23; следует, впрочем, учитывать, что ввиду комплексного характера гильбертова пространства луч этот только выглядит как обычная одномерная прямая, на деле же за ним скрывается целая комплексная плоскость.

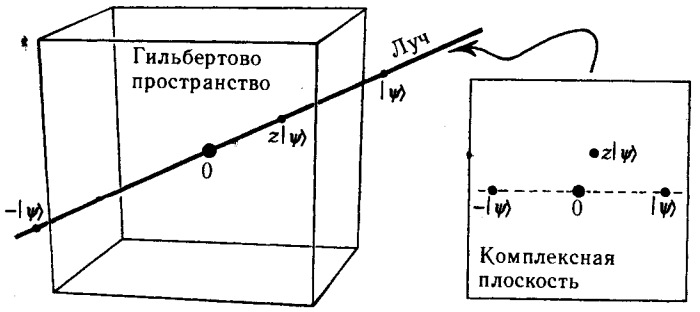



Рис. 5.23. *Луч* в гильбертовом пространстве есть множество всех комплексных кратных вектора состояния $|\psi\rangle$. Мы представляем этот луч в виде прямой, проходящей через начало гильбертовых координат, однако не следует забывать о том, что за этой прямой на деле скрывается комплексная плоскость.

До сих пор мы рассматривали гильбертово пространство, имея в виду лишь то, что структурно оно представляет собой комплексное векторное пространство. Однако, помимо комплексно-векторной структуры, у гильбертова пространства имеется еще одно, не менее важное, свойство, крайне полезное для описания процедуры редукции **R**. Речь идет об *эрмитовом скалярном произведении* (или *внутреннем произведении*), каковая операция позволяет из любой пары гильбертовых векторов получить одно-единственное комплексное число. Она же дает нам возможность ввести два весьма важных понятия. Первое — *квадрат длины* гильбертова вектора как скалярное произведение вектора на *самого себя*. Например, *нормированное* состояние (необходимое, как мы отмечали выше — см. § 5.8, с. 412, — для строгой применимости правила квадратов модулей) задается гильбертовым вектором, квадрат длины которого равен *единице*. Вторым важным понятием, сопутствующим скалярному произведению,



является понятие *ортогональности* гильбертовых векторов — векторы ортогональны, когда их скалярное произведение равно *нулю*. Ортогональными считаются векторы, направленные, в том или ином смысле, «под прямым углом» друг к другу. Применительно к состояниям, ортогональными обычно называют состояния, *независимые* одно от другого. Важность этого понятия для квантовой физики заключается в том, что различные альтернативные результаты любого измерения всегда ортогональны друг другу.

В качестве примера ортогональных состояний можно привести состояния $|\uparrow\rangle$ и $|\downarrow\rangle$, с которыми мы встречались при рассмотрении частицы со спином $\frac{1}{2}$. (Отметим, что ортогональность в гильбертовом пространстве, как правило, не соответствует перпендикулярности в пространстве обычном; в случае спина $\frac{1}{2}$ ортогональные состояния $|\uparrow\rangle$ и $|\downarrow\rangle$ представляют физические конфигурации, ориентированные, скорее, в противоположных направлениях, нежели под прямым углом.) Следующий пример — состояния $|\uparrow\uparrow\ \dots\ \uparrow\rangle, |\downarrow\uparrow\ \dots\ \uparrow\rangle, \dots, |\downarrow\downarrow\ \dots\ \downarrow\rangle$ спина $\frac{1}{2}n$; каждое такое состояние ортогонально всем остальным. Ортогональными являются и *все* различные возможные *положения*, в которых может находиться квантовая частица. Более того, ортогональны как состояния $|\mathbf{B}\rangle$ и $i|\mathbf{C}\rangle$ (см. § 5.7 — прошедшая и отраженная части состояния фотона, получаемые в результате падения фотона на полупрозрачное зеркало), так и состояния $i|\mathbf{D}\rangle$ и $-|\mathbf{E}\rangle$, в которые эволюционируют первые два после отражения от двух непрозрачных зеркал.

Последний факт иллюстрирует одно важное свойство шрёдингеровой эволюции **U**. Любые два изначально ортогональных состояния ортогональными и остаются, если каждое эволюционирует в соответствии с **U** в течение одного и того же временно́го периода. Таким образом, свойство ортогональности при эволюции **U** *сохраняется*. Кроме того, эволюция **U** сохраняет и *значение* скалярного произведения состояний. Собственно, именно в этом и заключается формальный смысл понятия *унитарная эволюция*.

Как уже упоминалось выше, ключевая роль ортогональности состоит в следующем: различные возможные квантовые состояния, возникающие при любом «измерении» квантовой си-

стемы и дающие — при поднятии на классический уровень — *непосредственно различимые* результаты, непременно ортогональны друг другу. Особенно наглядно это проявляется в *нулевых* измерениях — таких, например, как в задаче об испытании бомб, §§ 5.2 и 5.9. *Не*-обнаружение какого-либо квантового состояния устройством, способным это состояние обнаружить, приводит в конечном счете к тому, что результирующее состояние «перескакивает» в нечто, *ортогонально* противоположное тому состоянию, какое детектор, собственно, призван обнаруживать.

Как мы только что отметили, ортогональность математически выражается как *обращение в нуль* скалярного произведения состояний. Это скалярное произведение, в общем случае, представляет собой комплексное число, поставленное в соответствие какой-либо паре элементов гильбертова пространства. Если обозначить эти элементы (или состояния) через $|\psi\rangle$ и $|\phi\rangle$, то упомянутое комплексное число записывается так: $\langle\psi|\phi\rangle$. При этом выполняется ряд простых алгебраических тождеств, которые мы можем записать в следующем (несколько, правда, неуклюжем) виде:

$$\overline{\langle\psi|\phi\rangle} = \langle\phi|\psi\rangle,$$
$$\langle\psi|(|\phi\rangle + |\chi\rangle) = \langle\psi|\phi\rangle + \langle\psi|\chi\rangle,$$
$$(z\langle\psi|)|\phi\rangle = z\langle\psi|\phi\rangle,$$
$$\langle\psi|\psi\rangle > 0, \quad \text{кроме случая } |\psi\rangle = \mathbf{0}.$$

Кроме того, можно показать, что $\langle\psi|\psi\rangle = 0$ при $|\psi\rangle = \mathbf{0}$. Мне не хочется надоедать читателю прочими математическими подробностями (если же таковые подробности кого-то заинтересуют, то ознакомиться с ними можно, открыв любой стандартный текст по квантовой теории; см., например, [94]).

Существенными для наших дальнейших нужд свойствами скалярного произведения являются лишь следующие два (уже, впрочем, упоминавшиеся выше):

векторы $|\psi\rangle$ и $|\phi\rangle$ *ортогональны* тогда и только тогда, когда $\langle\psi|\phi\rangle = 0$,
произведение $\langle\psi|\psi\rangle$ есть *квадрат длины* вектора $|\psi\rangle$.

Отметим, что отношение ортогональности является симметричным (поскольку $\overline{\langle\psi|\phi\rangle} = \langle\phi|\psi\rangle$). Более того, произведение $\langle\psi|\psi\rangle$ всегда представляет собой неотрицательное вещественное число,



из какового числа легко извлекается неотрицательный квадратный корень, который мы можем называть *длиной* (или величиной) вектора $|\psi\rangle$.

Поскольку при умножении любого вектора состояния на ненулевое комплексное число физическая интерпретация этого вектора никаких изменений не претерпевает, мы всегда можем *нормировать* состояние таким образом, чтобы длина соответствующего вектора стала равна единице, получив в результате так называемый *единичный вектор*, или *нормированное состояние*. Тут, впрочем, имеется некоторая неясность, так как мы можем умножить вектор состояния и на чистую фазу (число вида $e^{i\theta}$, где $\theta$ — вещественное число; см. § 5.10).

## 5.13. Описание редукции R в терминах гильбертова пространства

Как в терминах гильбертова пространства представить процедуру **R**? Рассмотрим простейший случай измерения (типа «да/нет»), при котором прибор делает запись **ДА** при достоверном обнаружении у измеряемого квантового объекта некоторого свойства и **НЕТ**, если обнаружить данное свойство не удается (или, что то же самое, прибор обнаруживает достоверное указание на то, что таким свойством измеряемый квантовый объект не обладает). Этот случай включает в себя и ту возможность, которая нас в настоящий момент как раз и интересует, — вариант **НЕТ** может оказаться *нулевым* измерением. Подобные измерения выполняют, например, детекторы фотонов из § 5.8. Они регистрируют результат **ДА**, обнаруживая прибытие фотона, и **НЕТ**, если обнаружения фотона не произошло. В данном случае измерение **НЕТ** является не чем иным, как нулевым измерением — измерением оно при этом быть не перестает, вследствие чего состояние системы «скачком» переходит в состояние, ортогональное тому, какое наблюдалось бы, получи мы при измерении результат **ДА**. Аналогичным образом, к нулевым можно непосредственно отнести и измерения спина (для атома со спином $\frac{1}{2}$) в опыте Штерна—Герлаха; можно говорить, что измерение дает результат **ДА**, если обнаруживается, что атом имеет спин $|\uparrow\rangle$ (что происходит, когда атом отклоняется в сторону, соответствующую направлению «вверх»), или **НЕТ**, если атом в эту сторону

не отклоняется, что дает нам спиновое состояние, ортогональное состоянию $|\uparrow\rangle$, т. е. $|\downarrow\rangle$.

Более сложные измерения всегда можно представить в виде последовательности измерений типа «да/нет». Рассмотрим, например, атом со спином $\frac{1}{2}n$. Чтобы не упустить ни одного из $n+1$ различных возможных результатов измерения доли спина, ориентированного в направлении «вверх», начнем с того, что зададим вопрос, не находится ли атом в спиновом состоянии, например, $|\uparrow\uparrow\ldots\uparrow\rangle$. Для ответа на вопрос попытаемся обнаружить атом в луче, соответствующем этому спиновому состоянию «единодушно вверх». Если измерение дает ответ **ДА**, то на этом наши мучения и заканчиваются. Если же мы получаем **НЕТ**, то измерение оказывается нулевым, и мы переходим к следующему вопросу: «Не находится ли атом в спиновом состоянии $|\downarrow\uparrow\ldots\uparrow\rangle$?» И так далее. Каждый раз ответ **НЕТ** следует считать нулевым измерением, каковое указывает лишь на то, что в данном случае не был получен ответ **ДА**. Запишем наши рассуждения более подробно. Предположим, что первоначально атом находится в спиновом состоянии

$$z_0|\uparrow\uparrow\uparrow\ldots\uparrow\rangle + z_1|\downarrow\uparrow\uparrow\ldots\uparrow\rangle + z_2|\downarrow\downarrow\uparrow\ldots\uparrow\rangle + \ldots + z_n|\downarrow\downarrow\downarrow\ldots\downarrow\rangle,$$

а мы выполняем измерение с целью выяснить, не ориентирован ли весь спин атома в направлении «вверх». Получив ответ **ДА**, мы удостоверяемся в том, что атом действительно находится в состоянии $|\uparrow\uparrow\uparrow\ldots\uparrow\rangle$, или, если точнее, «перескакивает» в состояние $|\uparrow\uparrow\uparrow\ldots\uparrow\rangle$ при измерении. Если же ответ **НЕТ**, то измерение является нулевым, и приходится предположить, что первоначальное состояние «перескакивает» в ортогональное состояние

$$z_1|\downarrow\uparrow\uparrow\ldots\uparrow\rangle + z_2|\downarrow\downarrow\uparrow\ldots\uparrow\rangle + \ldots + z_n|\downarrow\downarrow\downarrow\ldots\downarrow\rangle.$$

Мы выполняем следующее измерение, на этот раз желая выяснить не находится ли атом в состоянии $|\downarrow\uparrow\uparrow\ldots\uparrow\rangle$. Получив при *этом* измерении ответ **ДА**, мы говорим, что атом и в самом деле находится в состоянии $|\downarrow\uparrow\uparrow\ldots\uparrow\rangle$ или, что правильнее, «перескакивает» в состояние $|\downarrow\uparrow\uparrow\ldots\uparrow\rangle$ в результате измерения. Если же мы получаем ответ **НЕТ**, то происходит «скачок» в следующее состояние,

$$z_2|\downarrow\downarrow\uparrow\ldots\uparrow\rangle + \ldots + z_n|\downarrow\downarrow\downarrow\ldots\downarrow\rangle,$$

и так далее.



Эти «скачки», совершаемые (или, по крайней мере, *кажущиеся* совершаемыми) вектором состояния, олицетворяют собой наиболее головоломный аспект квантовой теории. Думаю, недалеко от истины утверждение, что большинство квантовых физиков либо испытывают немалые *трудности*, пытаясь примириться с тем фактом, что подобные «скачки» неотъемлемо присущи объективной физической реальности, либо *вообще* отказываются признавать, что реальность может вести себя столь абсурдным образом. Тем не менее, какой бы точки зрения относительно связи описываемых здесь процессов с «реальностью» мы ни придерживались, упомянутые «скачки» представляют собой существенный элемент квантового формализма.

В предыдущем рассуждении я воспользовался правилом, иногда называемым *проекционным постулатом* и однозначно определяющим форму подобных «скачков» (например, состояние $z_0|\uparrow\uparrow\ldots\uparrow\rangle + z_1|\downarrow\uparrow\ldots\uparrow\rangle + \ldots + z_n|\downarrow\downarrow\ldots\downarrow\rangle$ должно «перескакивать» в состояние $z_1|\downarrow\uparrow\ldots\uparrow\rangle + \ldots + z_n|\downarrow\downarrow\ldots\downarrow\rangle$). Название постулата обусловлено геометрическими соображениями, в чем мы вскоре убедимся. По мнению некоторых физиков, проекционный постулат представляет собой несущественное допущение квантовой теории. Физики эти, впрочем, имеют в виду, как правило, отнюдь не нулевые измерения, но измерения, при которых квантовое состояние *нарушается* неким физическим взаимодействием. Такое нарушение происходит, когда измерение (в вышеописанных примерах) дает ответ **ДА**, т. е. детектор регистрирует фотон, поглощая его при этом, а атом по прохождении установки Штерна—Герлаха оказывается в некотором конкретном луче (что опять же означает **ДА**). Для рассматриваемого же нулевого измерения (т. е. измерения, при котором мы получаем ответ **НЕТ**) проекционный постулат оказывается как нельзя более существенным, поскольку без него никак невозможно узнать, что квантовая теория думает (и, кстати, правильно думает) по поводу измерений, следующих за нулевым.

Для того, чтобы получить более наглядное представление о смысле проекционного постулата, попробуем описать происходящее в терминах гильбертова пространства. Для этого введем понятие *примитивного* измерения. Примитивным я буду называть измерение типа «да/нет», при котором результат **ДА** означает, что система находится в некотором определенном квантовом состоянии $|\alpha\rangle$ (либо в кратном ему состоянии $u|\alpha\rangle$,

где $u \neq 0$) — или только что в это состояние «перескочила». Таким образом, в случае примитивного измерения результат **ДА** определяет физическое состояние системы как нечто конкретное и *единственное*, тогда как результат **НЕТ** может предполагать несколько альтернативных вариантов развития событий. Примитивными являются, например, описанные выше измерения спина, посредством которых мы пытались установить, не находится ли спин в том или ином состоянии (скажем, в состоянии $|\downarrow\downarrow\uparrow \ldots \uparrow\rangle$).

При примитивном измерении результат **НЕТ** *проецирует* состояние системы на состояние, ортогональное $|\alpha\rangle$. На рис. 5.24 представлена геометрическая интерпретация этой процедуры. За начальное состояние примем состояние $|\psi\rangle$ (обозначенное на рисунке большой стрелкой) — в результате измерения оно «перескакивает» либо в состояние, кратное $|\alpha\rangle$ (если ответ **ДА**), либо проецируется на состояние, ортогональное $|\alpha\rangle$ (если ответ **НЕТ**). Со случаем **НЕТ** никаких дополнительных проблем не возникает — согласно стандартной квантовой теории, именно такого результата и следует ожидать. В случае же ответа **ДА** ситуация осложняется тем, что здесь квантовая система вступает во взаимодействие с измерительным устройством, переходя в состояние, значительно более хитроумное, нежели просто $|\alpha\rangle$. Результатом такой эволюции оказывается, в общем случае, так называемое *сцепленное состояние*, «сплетающее» в одно целое исходную квантовую систему и измерительное устройство. (Сцепленные состояния мы рассмотрим в § 5.17.) Тем не менее, дальше квантовая система должна эволюционировать так, *будто* она и в самом деле перескочила в состояние, кратное $|\alpha\rangle$; в противном случае последующая эволюция системы становится неоднозначной.

Алгебраически этот скачок выражается следующим образом. Вектор состояния $|\psi\rangle$ всегда можно записать (в данном случае — однозначно, поскольку вектор $|\alpha\rangle$ задан) в виде

$$|\psi\rangle = z|\alpha\rangle + |\chi\rangle,$$

где $|\chi\rangle$ ортогонален $|\alpha\rangle$. Вектор $z|\alpha\rangle$ есть ортогональная проекция вектора $|\psi\rangle$ на луч, содержащий вектор $|\alpha\rangle$, а $|\chi\rangle$ — это ортогональная проекция $|\psi\rangle$ на *пространство ортогональных дополнений* $|\alpha\rangle$ (т. е. на пространство всех векторов, ортогональных $|\alpha\rangle$). Если измерение дает результат **ДА**, то это нужно понимать так, что вектор состояния перескочил в $z|\alpha\rangle$ (или просто



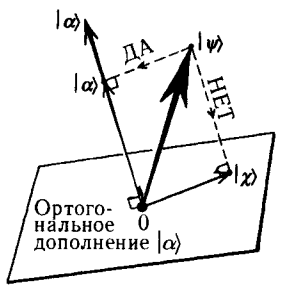



Рис. 5.24. Примитивное измерение проецирует состояние $|\psi\rangle$ в состояние, кратное заданному состоянию $|\alpha\rangle$ (в случае ответа **ДА**), или в состояние, являющееся ортогональным дополнением $|\alpha\rangle$ (в случае ответа **НЕТ**).

в $|\alpha\rangle$), что является отправной точкой его последующей эволюции. Если же результат **НЕТ**, то вектор перескакивает в $|\chi\rangle$.

Какие вероятности следует приписать каждому из двух альтернативных результатов? Для того, чтобы воспользоваться предложенным выше «правилом квадратов модулей», будем полагать вектор $|\alpha\rangle$ *единичным* и выберем некоторый единичный вектор $|\phi\rangle$ в направлении вектора $|\chi\rangle$, т. е. $|\chi\rangle = w|\phi\rangle$. Тогда выражение принимает вид

$$|\psi\rangle = z|\alpha\rangle + w|\phi\rangle$$

(где, собственно, $z = \langle\alpha|\psi\rangle$ и $w = \langle\phi|\psi\rangle$), а относительные вероятности результатов **ДА** и **НЕТ** вычисляются через отношение квадратов $|z|^2$ и $|w|^2$. Если и сам вектор $|\psi\rangle$ является единичным, то величины $|z|^2$ и $|w|^2$ представляют собой *фактические* вероятности, соответственно, результатов **ДА** и **НЕТ**.

Можно сформулировать все это и по-другому, причем в настоящем контексте получится даже несколько проще (в качестве упражнения предлагаю заинтересованному читателю самостоятельно убедиться в том, что эти формулировки эквивалентны). Для того чтобы определить фактическую вероятность каждого из возможных результатов (в данном случае, **ДА** и **НЕТ**), мы просто возводим в квадрат длину вектора $|\psi\rangle$ (ненормированного к единичному вектору), после чего сравниваем полученное значение с квадратами длины соответствующих проекций. Коэффициент

уменьшения в каждом случае и будет представлять собой искомую вероятность.

В заключение следует упомянуть, что в случае *общего* измерения типа «да/нет» (т. е. не только примитивного), когда **ДА**-состояния не обязательно принадлежат одному-единственному лучу, рассуждение будет по большей части аналогично вышеприведенному. Только здесь речь пойдет о **ДА**-подпространстве **Д** и **НЕТ**-подпространстве **Н**. Эти подпространства являются ортогональными дополнениями друг друга — в том смысле, что любой вектор одного ортогонален любому вектору другого, вместе же они заполняют все исходное гильбертово пространство. Согласно проекционному постулату, при измерении первоначальный вектор состояния $|\psi\rangle$ ортогонально проецируется на подпространство **Д**, если получен ответ **ДА**, и на подпространство **Н**, если получен ответ **НЕТ**. Относительные вероятности этих результатов здесь также определяются коэффициентами уменьшения квадрата длины вектора состояния при соответствующем проецировании (см. НРК, с. 263, рис. 6.23). Впрочем, статус проекционного постулата в данном случае представляется несколько менее ясным, чем при нулевом измерении, поскольку при утвердительном результате измерения результирующее состояние сцепляется с состоянием измерительного устройства. Поэтому в последующих рассуждениях я ограничусь более простыми *примитивными* измерениями, **ДА**-пространство которых состоит из одного-единственного луча (содержащего векторы, кратные $|\psi\rangle$). Для наших нужд этого будет вполне достаточно.

## 5.14. Коммутирующие измерения

При проведении нескольких последовательных измерений квантовой системы порядок, в котором эти измерения выполняются, может быть, в общем случае, важным. Измерения, от порядка выполнения которых зависит, какой вектор состояния мы получим в конечном итоге, называются *некоммутирующими*. Если же порядок выполнения измерений не играет абсолютно никакой роли (не изменяется даже фаза результирующего состояния), то мы говорим, что такие измерения *коммутируют*. В терминах гильбертова пространства это можно понимать так: при нескольких последовательных ортогональных проекциях заданного вектора состояния $|\psi\rangle$ окончательный результат, как



правило, зависит от порядка выполнения этих проекций. В случае коммутирующих измерений порядок их выполнения никакой роли не играет.

Что же происходит в случае *примитивных* измерений? Нетрудно убедиться, что для коммутируемости двух различных примитивных измерений необходимо, чтобы **ДА**-луч одного был *ортогонален* **ДА**-лучу другого.

Например, примитивные измерения спина атома со спином $\frac{1}{2}n$ (см. § 5.10) можно выполнять в любом порядке, так как все возможные состояния здесь ($|\uparrow\uparrow\ldots\uparrow\rangle$, $|\downarrow\uparrow\ldots\uparrow\rangle$, …, $|\downarrow\downarrow\ldots\downarrow\rangle$) ортогональны друг другу. Таким образом, окончательный результат измерения никак не зависит от выбранного мной конкретного порядка выполнения примитивных измерений — все эти измерения коммутируют. Впрочем, в общем случае это не всегда так — например, нам может вздуматься выполнять отдельные измерения спина относительно различных направлений. *Такие* измерения, как правило, не коммутируют.

## 5.15. Квантовомеханическое «И»

В квантовой механике имеется стандартная процедура для исследования систем из двух и более независимых компонентов. Эта процедура понадобится нам, в частности, при рассмотрении с квантовой точки зрения (которое мы планируем дать в § 5.18) системы, состоящей из двух далеко разнесенных в пространстве частиц со спином $\frac{3}{2}$ — тех самых частиц, которые «Квинтэссенциальные Товары» поместили в магические додекаэдры (см. § 5.3). Необходима такая процедура и для квантовомеханического описания детектора в момент сцепления его состояния с квантовым состоянием регистрируемой частицы.

Рассмотрим для начала систему, состоящую всего из *двух* независимых (невзаимодействующих) компонентов. Допустим, что каждый из этих компонентов (в отсутствие другого) описывается своим вектором состояния — скажем, $|\alpha\rangle$ и $|\beta\rangle$. Как описать *всю* систему, в которой присутствуют *оба* компонента? Обычная процедура заключается в составлении так называемого *тензорного* (или *внешнего*) *произведения* этих векторов, которое записывается следующим образом:

$$|\alpha\rangle|\beta\rangle.$$

Мы можем рассматривать это произведение как стандартный квантовомеханический способ представления обыкновенного логического «И» — в том смысле, что такая система объединяет в себе в некоторый момент времени *обе* независимые квантовые системы, представленные, соответственно, векторами состояния $|\alpha\rangle$ и $|\beta\rangle$. (Например, $|\alpha\rangle$ может представлять электрон, находящийся в точке A, а $|\beta\rangle$ — атом водорода в некоторой отдаленной точке B. Тогда состояние, в котором электрон находится в точке A, а атом водорода — в точке B, будет представлено произведением $|\alpha\rangle|\beta\rangle$.) Величина $|\alpha\rangle|\beta\rangle$ представляет *одно* квантовое состояние — мы вполне можем обозначить его одним вектором состояния, скажем, $|\chi\rangle$, и, не нарушив ни одного закона, записать

$$|\chi\rangle = |\alpha\rangle|\beta\rangle.$$

Следует особо подчеркнуть, что это понятие «И» не имеет ничего общего с квантовой линейной суперпозицией, которая записывается как сумма векторов состояний $|\alpha\rangle + |\beta\rangle$ или, в общем случае, $z|\alpha\rangle + w|\beta\rangle$, где $z$ и $w$ — комплексные весовые коэффициенты. Например, если $|\alpha\rangle$ и $|\beta\rangle$ — возможные состояния одного фотона (соответствующие, скажем, его расположению в различных точках A и B), то запись $|\alpha\rangle + |\beta\rangle$ также представляет возможное состояние *того же самого* фотона, при котором он замирает в нерешительности где-то между A и B в соответствии с маловразумительными предписаниями квантовой теории, — *одного* фотона, заметим, никак не *двух*. Состояние *пары* фотонов, при котором один находится в точке A, а другой — в точке B, будет представлено уже вектором $|\alpha\rangle|\beta\rangle$.

Тензорное произведение подчиняется тем же алгебраическим правилам, каким, по нашим представлениям, и должно подчиняться любое уважающее себя произведение:

$$\begin{aligned}(z|\alpha\rangle)|\beta\rangle &= z(|\alpha\rangle|\beta\rangle) = |\alpha\rangle(z|\beta\rangle),\\ (|\alpha\rangle + |\gamma\rangle)|\beta\rangle &= |\alpha\rangle|\beta\rangle + |\gamma\rangle|\beta\rangle,\\ |\alpha\rangle(|\beta\rangle + |\gamma\rangle) &= |\alpha\rangle|\beta\rangle + |\alpha\rangle|\gamma\rangle,\\ (|\alpha\rangle|\beta\rangle)|\gamma\rangle &= |\alpha\rangle(|\beta\rangle|\gamma\rangle),\end{aligned}$$

разве что равенство $|\alpha\rangle|\beta\rangle = |\beta\rangle|\alpha\rangle$, строго говоря, некорректно. Это, впрочем, отнюдь не означает, что интерпретация понятия «И» в квантовомеханическом контексте предполагает, что совокупная система «$|\alpha\rangle$ и $|\beta\rangle$» физически чем-то отличается от совокупной системы «$|\beta\rangle$ и $|\alpha\rangle$». Мы попробуем обойти эту проблему посредством несколько более глубокого погружения в таинства действительного поведения Вселенной на квантовом уровне. В дальнейшем под записью $|\alpha\rangle|\beta\rangle$ мы будем подразумевать не то, что математики называют «тензорным произведением», а скорее то, что в математической физике (с недавних пор) называется *грассмановым произведением*. Тогда к записанным выше можно добавить еще одно правило:



$$|\beta\rangle|\alpha\rangle = \pm|\alpha\rangle|\beta\rangle.$$

Знак «минус» появляется здесь лишь в том случае, когда *оба* состояния ($|\alpha\rangle$ и $|\beta\rangle$) «охватывают» нечетное количество частиц с нецелочисленным спином. (Такие частицы называются *фермионами*, а их спин принимает значения $\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \frac{7}{2}, \ldots$. Частицы со спином $0, 1, 2, 3, \ldots$ называются *бозонами* и на знак в приведенном выше выражении никак не влияют.) Впрочем, на данном этапе читателю нет необходимости вникать во все эти формальности. До тех пор, пока нас занимает лишь скрывающееся за описанием физическое состояние, «$|\alpha\rangle$ и $|\beta\rangle$» ничем не отличается от «$|\beta\rangle$ и $|\alpha\rangle$».

Для описания состояний с тремя или большим количеством независимых компонентов мы просто повторяем процедуру. Так, если обозначить индивидуальные состояния этих трех компонентов через $|\alpha\rangle$, $|\beta\rangle$ и $|\gamma\rangle$, то состояние, в котором все три компонента наличествуют одновременно, описывается произведением

$$|\alpha\rangle|\beta\rangle|\gamma\rangle,$$

причем грассманово произведение $(|\alpha\rangle|\beta\rangle)|\gamma\rangle$ (или, что эквивалентно, $|\alpha\rangle(|\beta\rangle|\gamma\rangle)$) описывает то же самое состояние. Аналогичным образом рассматриваются и системы с четырьмя или более независимыми компонентами.

Следует упомянуть и об одном важном свойстве шрёдингеровой эволюции **U**: эволюция совокупной системы $|\alpha\rangle|\beta\rangle$ (где $|\alpha\rangle$ и $|\beta\rangle$ никак друг с другом не взаимодействуют) есть не что иное, как совокупность эволюций индивидуальных систем. Так, если по истечении некоторого времени $t$ система $|\alpha\rangle$ эволюционирует (индивидуально) в систему $|\alpha'\rangle$, а система $|\beta\rangle$ эволюционирует

(индивидуально) в систему $|\beta'\rangle$, то совокупная система $|\alpha\rangle|\beta\rangle$ за то же время $t$ эволюционирует в систему $|\alpha'\rangle|\beta'\rangle$. Аналогично, если у нас имеется три невзаимодействующих компонента $|\alpha\rangle$, $|\beta\rangle$ и $|\gamma\rangle$, эволюционирующих, соответственно, в $|\alpha'\rangle$, $|\beta'\rangle$ и $|\gamma'\rangle$, то совокупная система $|\alpha\rangle|\beta\rangle|\gamma\rangle$ посредством той же эволюции переходит в состояние $|\alpha'\rangle|\beta'\rangle|\gamma'\rangle$. То же верно для четырех и более компонент.

Отметим, что свойство это очень похоже на свойство *линейности* эволюции **U** (см. §5.7), согласно которому результат эволюции суперпозиции состояний в точности совпадает с суперпозицией результатов эволюции отдельных состояний. Состояние $|\alpha\rangle+|\beta\rangle$, например, эволюционирует в $|\alpha'\rangle+|\beta'\rangle$. Тем не менее, речь в обоих случаях идет о совершенно *разных* вещах, и очень важно об этой разнице не забывать. Нет ничего удивительного в том, что система, составленная из невзаимодействующих независимых компонентов, эволюционирует — как целое — так, словно ни один из ее отдельных компонентов понятия не имеет о присутствии в системе остальных. Независимость компонентов (т. е. полное отсутствие каких бы то ни было взаимодействий между ними) в данном случае — существенное условие, иначе свойство не «работает». Свойство линейности же оказывается поистине неожиданным. Получается, что под действием **U** системы-суперпозиции состояний эволюционируют как набор отдельных, полностью изолированных друг от друга состояний *независимо* от того, изолированы эти состояния в действительности или между ними существуют какие-то взаимодействия. Одного этого достаточно, чтобы усомниться в абсолютной справедливости свойства линейности. И все же эволюция **U** линейна (и тому есть многочисленные подтверждения), но лишь в отношении феноменов, целиком и полностью ограниченных квантовым уровнем. Нарушение же линейности происходит, по всей видимости, исключительно под действием процедуры **R**. К этому вопросу мы еще вернемся.

## 5.16. Ортогональность произведений состояний

С ортогональностью произведений состояний (в том виде, в каком я определил эти произведения выше) дела обстоят не так просто, как хотелось бы. Допустим, у нас имеется два *ортогональных состояния* $|\alpha\rangle$ и $|\beta\rangle$; тогда мы вправе ожидать, что состояния $|\psi\rangle|\alpha\rangle$ и $|\psi\rangle|\beta\rangle$ также будут ортогональными, причем при любом $|\psi\rangle$. Пусть, например, $|\alpha\rangle$ и $|\beta\rangle$ — возможные альтернативные состояния фотона, где $|\alpha\rangle$ — состояние фотона, зарегистрированного неким фотоэлементом, а ортогональное $|\alpha\rangle$ состояние $|\beta\rangle$ — *предполагаемое* состояние фотона в случае, когда фотоэлемент не регистрирует ничего (нулевое измерение). Можно представить себе, что наш фотон является компонентом некоей совокупной системы — просто добавим к нему еще какой-нибудь объект (например, другой фотон, скажем, где-нибудь на Луне) и обозначим состояние этого другого объекта через $|\psi\rangle$. Таким образом, для нашей совокупной системы возможны два альтернативных состояния — $|\psi\rangle|\alpha\rangle$ и $|\psi\rangle|\beta\rangle$. Простое добавление состояния $|\psi\rangle$ в имеющееся описание не должно, разумеется, оказать никакого влияния на ортогональность двух первоначальных состояний. В самом деле, если говорить об определении произведения состояний в терминах обычного «тензорного произведения» (или необычного — в данном случае, грассманова произведения, а точнее, некоторой его модификации, используемой в наших рассуждениях), то так оно и есть, и из ортогональности состояний $|\alpha\rangle$ и $|\beta\rangle$ действительно следует ортогональность $|\psi\rangle|\alpha\rangle$ и $|\psi\rangle|\beta\rangle$.

Как бы то ни было, пути, которыми, похоже (согласно последним данным квантовой теории), предпочитает следовать Вселенная, далеко не столь прямолинейны. Если бы состояние $|\psi\rangle$ можно было счесть полностью независимым и от $|\alpha\rangle$, и от $|\beta\rangle$, то тогда его присутствие и в самом деле ничего бы не меняло. Однако формально полной независимости здесь быть не может, и состояние даже пребывающего на Луне фотона оказывает самое непосредственное воздействие на состояние фотона, регистрируемого нашим фотоэлементом[9]. (С этими формальностями связано, в частности, то, что под обозначением «$|\psi\rangle|\alpha\rangle$» мы подразумеваем произведение грассманова типа — если использовать более привычные термины, то речь тут идет о так называемой

[9] Любопытно, что такого рода феномены находят недвусмысленное подтверждение в реальных физических наблюдениях. Описанный Хэнбери Брауном и Твиссом [187, 188] эффект, в соответствии с которым были измерены диаметры некоторых близлежащих звезд, основывается как раз на таком «бозонном» свойстве взаимодействия достигающих Земли фотонов, испущенных с противоположных краев звезды.

«статистике Бозе» (описание состояний фотонов и прочих бозонов) или о «статистике Ферми» (описание состояний фермионов — электронов, протонов и т. д.), см. НРК, с. 277, 278 и, скажем, [94].) Если бы перед нами стояла задача получить абсолютно точный с точки зрения теории результат, то рассмотрение состояния одного-единственного фотона потребовало бы учета состояний всех фотонов во Вселенной. Впрочем, необходимости в этом (к счастью) нет — и без такого учета точность получаемых результатов хоть и не абсолютна, но все же чрезвычайно высока. Если состояния $|\alpha\rangle$ и $|\beta\rangle$ ортогональны, то можно с очень высокой степенью точности предположить, что ортогональными будут и состояния $|\psi\rangle|\alpha\rangle$ и $|\psi\rangle|\beta\rangle$ (даже если это произведения грассманова типа), где $|\psi\rangle$ — любое состояние, не имеющее очевидного отношения к рассматриваемой задаче (каковая задача непосредственно касается лишь ортогональных состояний $|\alpha\rangle$ и $|\beta\rangle$). Так и предположим.

## 5.17. Квантовая сцепленность

Для того чтобы двигаться дальше, нам не обойтись без понимания квантовой физики *ЭПР-эффектов* — квантовомеханических **Z**-загадок, ярким представителем которых является представленная мною выше задача о магических додекаэдрах (см. §§ 5.3, 5.4). Кроме того, мы должны как-то разобраться с главной **X**-загадкой квантовой теории — парадоксальной взаимозависимостью между процессами эволюции **U** и редукции **R**, загадкой, порождающей *проблему измерения*, о которой мы поговорим в следующей главе. Следовательно, настала пора ввести очередную фундаментальную квантовую идею — понятие о *сцепленных состояниях*.

Начнем с того, что попытаемся выяснить, что включает в себя простой процесс измерения. Рассмотрим следующую ситуацию: фотон находится в суперпозиции, скажем, $|\alpha\rangle + |\beta\rangle$, где в состоянии $|\alpha\rangle$ фотон активирует детектор, в состоянии же $|\beta\rangle$, ортогональном $|\alpha\rangle$, фотон никакого воздействия на детектор не оказывает. (Похожий пример рассматривался в § 5.8, когда на детектор, расположенный в точке **G**, падал фотон, пребывающий в состоянии $-|\mathbf{F}\rangle - i|\mathbf{G}\rangle$. В состоянии $|\mathbf{G}\rangle$ фотон активировал детектор, в состоянии $|\mathbf{F}\rangle$ никакого воздействия на детектор не



происходило.) Предположим далее, что детектору тоже можно сопоставить некое квантовое состояние, скажем, $|\Psi\rangle$. Вообще говоря, в квантовой теории это обычная практика. Лично мне не совсем ясно, какой может быть смысл в придании квантовомеханического описания объекту классического уровня, однако в дискуссиях на эту тему подобные вопросы, как правило, никого не занимают. Как бы то ни было, мы, думаю, можем согласиться с тем, что те элементы детектора, с которыми фотон сталкивается *прежде всего*, и в самом деле допускают рассмотрение согласно стандартным правилам квантовой теории. Поэтому, если у вас возникают какие-либо сомнения относительно правомерности применения этих правил ко всему детектору (как к целому), вы можете считать, что вектор состояния $|\Psi\rangle$ описывает поведение именно совокупности элементов квантового уровня (частиц, атомов, молекул), что принимают на себя, так сказать, первый удар.

В момент, непосредственно предшествующий столкновению фотона (или, точнее, $|\alpha\rangle$-части волновой функции фотона) с детектором, физическое состояние системы объединяет в себе состояние детектора *и* состояние фотона, т. е. имеет вид $|\Psi\rangle(|\alpha\rangle + |\beta\rangle)$, а нам известно, что

$$|\Psi\rangle(|\alpha\rangle + |\beta\rangle) = |\Psi\rangle|\alpha\rangle + |\Psi\rangle|\beta\rangle.$$

Таким образом, мы имеем дело с суперпозицией состояния $|\Psi\rangle|\alpha\rangle$, описывающего детектор (элементы детектора) и приближающийся к нему фотон, и состояния $|\Psi\rangle|\beta\rangle$, описывающего детектор (элементы детектора) и фотон, находящийся где-то в другом месте. Предположим далее, что состояние $|\Psi\rangle|\alpha\rangle$ (детектор с приближающимся к нему фотоном) переходит, согласно шрёдингеровой эволюции **U**, в некоторое новое состояние $|\Psi_{\text{Д}}\rangle$ (детектор регистрирует результат **ДА**) — в силу возникающих при столкновении взаимодействий между фотоном и элементами детектора. Предположим также, что если фотон с детектором *не* сталкивается, то под действием **U** состояние детектора $|\Psi\rangle$ эволюционирует (индивидуально) в состояние $|\Psi_{\text{Н}}\rangle$ (детектор регистрирует **НЕТ**), а состояние $|\beta\rangle$ — в состояние $|\beta'\rangle$. Тогда, согласно свойствам шрёдингеровой эволюции, рассмотренным в предыдущем параграфе, общее состояние системы принимает вид

$$|\Psi_{\text{Д}}\rangle + |\Psi_{\text{Н}}\rangle|\beta'\rangle.$$

29*

Это аккуратнее, чем обычно
Всос. $|\text{Д}\rangle$ фотон может исчезнуть

Перед нами типичный пример *сцепленного* состояния: термин «сцепленность» в данном случае отражает тот факт, что общее состояние системы невозможно записать просто в виде *произведения* состояния одной из ее подсистем (фотона) на состояние другой подсистемы (детектора). Более того, состояние $|\Psi_Д\rangle$ и само, по всей вероятности, является сцепленным (по меньшей мере с состояниями элементов собственного окружения), однако подтверждение этой сцепленности требует детального исследования соответствующих взаимодействий, не имеющих к теме нашего разговора никакого отношения.

Отметим, что состояния $|\Psi\rangle|\alpha\rangle$ и $|\Psi\rangle|\beta\rangle$, суперпозицией которых представлено состояние совокупной системы непосредственно перед столкновением, (существенно) *ортогональны* — поскольку ортогональны состояния $|\alpha\rangle$ и $|\beta\rangle$, а $|\Psi\rangle$ никак не зависит ни от того, ни от другого. Таким образом, ортогональными должны быть и состояния, в которые они эволюционируют под действием **U**, — $|\Psi_Д\rangle$ и $|\Psi_Н\rangle|\beta'\rangle$. (Эволюция **U** всегда сохраняет ортогональность.) Состояние $|\Psi_Д\rangle$ может в дальнейшем эволюционировать в нечто, наблюдаемое на макроскопическом уровне, — например, в слышимый человеческим ухом щелчок, указывающий на то, что фотон действительно был зарегистрирован. Если же никакого щелчка мы не услышали, то это надо понимать так, что система находится в ортогональном альтернативном состоянии $|\Psi_Н\rangle|\beta'\rangle$ (или только что в него «перескочила»). Одна лишь контрфактуальная возможность — щелчок *мог* прозвучать, но не прозвучал — вызывает «скачок» состояния из суперпозиции в состояние $|\Psi_Н\rangle|\beta'\rangle$, причем новое состояние уже *не* является сцепленным. Его *рас*цепило нулевое измерение.

Характерной особенностью сцепленных состояний является то, что «скачок», сопровождающий процедуру **R**, может в данном случае иметь, на первый взгляд, нелокальное (или даже явно ретроактивное) действие, еще более удивительное, чем результат простого нулевого измерения. Такая нелокальность, в частности, имеет место в так называемых ЭПР-эффектах (или феноменах Эйнштейна — Подольского — Розена). Эти эффекты — подлинные квантовые чудеса — можно отнести к наиболее непостижимым **Z**-загадкам квантовой теории. Идею подобного парадокса первоначально выдвинул Эйнштейн, желая показать, что формализм квантовой теории не в состоянии дать исчерпывающее описание Вселенной. Впоследствии было предложено множество



различных вариантов ЭПР-феноменов (например, магические додекаэдры из § 5.3), причем некоторые из них получили прямое экспериментальное подтверждение, т. е. оказались неотъемлемой частью *действительного* устройства мира, в котором мы живем (см. § 5.4).

ЭПР-эффекты возникают в следующего рода ситуациях. Рассмотрим известное начальное состояние $|\Omega\rangle$ физической системы, которое эволюционирует (согласно **U**) в суперпозицию двух ортогональных состояний, каждое из которых представляет собой произведение двух независимых состояний, описывающих два пространственно разделенных физических компонента системы — т. е. $|\Omega\rangle$ эволюционирует, скажем, в сцепленное состояние

$$|\psi\rangle|\alpha\rangle + |\phi\rangle|\beta\rangle.$$

Допустим, состояния $|\psi\rangle$ и $|\phi\rangle$ — это ортогональные альтернативы для одного компонента системы, а $|\alpha\rangle$ и $|\beta\rangle$ — ортогональные альтернативы для другого компонента. Измерение, устанавливающее в каком из состояний, $|\psi\rangle$ или $|\phi\rangle$, находится первый компонент, тем самым немедленно определяет и соответствующее состояние ( $|\alpha\rangle$ или $|\beta\rangle$) второго компонента.

Пока, кажется, ничего сверхъестественного. Кто-то может даже предположить, что нечто очень похожее мы могли наблюдать в случае с добрым доктором Бертлманом и его носками (§ 5.4). Коль скоро нам известно, что носки доктора должны быть разного цвета, — и кроме того, мы выяснили, что сегодня он остановил свой выбор, скажем, на зеленом и розовом, — то наблюдение, устанавливающее, что левый носок доктора зеленый (состояние $|\psi\rangle$) или же розовый (состояние $|\phi\rangle$), немедленно определяет цвет его правого носка — соответственно, розового (состояние $|\alpha\rangle$) или зеленого (состояние $|\beta\rangle$). Как бы то ни было, эффекты квантовой сцепленности могут фундаментально отличаться от вышеописанного, и никакая «бертлмано-носочная» трактовка не в состоянии объяснить все наблюдаемые результаты. Серьезные проблемы начинаются тогда, когда компоненты системы могут быть измерены несколькими *альтернативными* способами.

Проиллюстрируем сказанное примером. Предположим, что начальное состояние $|\Omega_0\rangle$ описывает спиновое состояние некоторой частицы как спин 0. Частица затем распадается на две новые

частицы (каждая со спином $\frac{1}{2}$), которые разлетаются в разные стороны (скажем, влево и вправо), удаляясь на значительное расстояние друг от друга. Из свойств кинетического момента и из закона его сохранения следует, что спины образовавшихся при распаде частиц должны быть ориентированы в противоположном направлении; таким образом, состояние нулевого спина, в которое эволюционирует $|\Omega_0\rangle$, имеет вид

$$|\Omega\rangle = |\mathbf{L}\uparrow\rangle|\mathbf{R}\downarrow\rangle - |\mathbf{L}\downarrow\rangle|\mathbf{R}\uparrow\rangle,$$

где «**L**» обозначает частицу, движущуюся влево, а «**R**» — частицу, движущуюся вправо (знак «минус» появляется согласно стандартному правилу). Допустим, мы решаем провести измерение спина левой частицы на предмет направленности его оси «вверх». Тогда ответ **ДА** (т. е. обнаружение состояния $|\mathbf{L}\uparrow\rangle$) автоматически поместит правую частицу в состояние $|\mathbf{R}\downarrow\rangle$ («спин вниз»). Ответ **НЕТ** ($|\mathbf{L}\downarrow\rangle$) автоматически помещает правую частицу в состояние «спин вверх» ($|\mathbf{R}\uparrow\rangle$). Похоже, что измерение частицы «здесь» способно мгновенно повлиять на состояние частицы «там» (причем это «там» может быть очень далеко отсюда) — что, впрочем, ничуть не более удивительно, чем все те же «бертлмановские носки»!

Однако это сцепленное состояние можно представить и иначе, для этого нужно всего лишь выполнить другое измерение. Например, мы могли бы выбрать при измерении спина левой частицы другое направление — не вертикальное, а *горизонтальное*, т. е. ответ **ДА** соответствовал бы состоянию, скажем, $|\mathbf{L}\leftarrow\rangle$, а ответ **НЕТ** — состоянию $|\mathbf{L}\rightarrow\rangle$. Путем простого вычисления (см. НРК, с. 283) находим, что *то же* совокупное состояние $|\Omega\rangle$ можно записать иначе:

$$|\Omega\rangle = |\mathbf{L}\leftarrow\rangle|\mathbf{R}\rightarrow\rangle - |\mathbf{L}\rightarrow\rangle|\mathbf{R}\leftarrow\rangle.$$

Таким образом, ответ **ДА** при измерении левой частицы автоматически помещает правую частицу в состояние $|\mathbf{R}\rightarrow\rangle$, а ответ **НЕТ** — в состояние $|\mathbf{R}\leftarrow\rangle$. *Какое бы направление* для измерения спина левой частицы мы ни выбрали, мы получим соответствующий, отличный от прочих, результат.

Что в подобного рода ситуациях замечательно, так это то, что простой *выбор* направления оси спина левой частицы *определяет*, судя по всему, направление оси спина правой частицы. Более того, пока не получен *результат* левого измерения, никакой



реальной информации правой частице не передается. Одно лишь «установление направления оси спина» не производит, само по себе, никакого реально наблюдаемого эффекта. Несмотря на то, что сегодня все это хорошо понимают, до сих пор встречаются люди, которые тешат себя надеждой отыскать способ использовать ЭПР-эффект для *мгновенной* передачи сигналов из одного места в другое, ведь редукция вектора состояния **R** «редуцирует» квантовое состояние ЭПР-пары частиц мгновенно, вне зависимости от того, какое расстояние их разделяет. Как это ни печально, однако способа передать посредством описанной процедуры сигнал от левой частицы к правой не существует (см. [145]).

Согласно стандартному квантовомеханическому формализму все, действительно, так и выглядит: немедленно по выполнении измерения, скажем, левой частицы происходит редукция полного состояния системы — из начального сцепленного состояния (где ни одна частица *в отдельности* определенного спинового состояния не имеет) в состояние, при котором левое состояние «расцепляется» с правым, а оба спина приобретают вполне определенное значение. В *математическом* описании в терминах вектора состояния измерение слева и в самом производит на правую частицу мгновенное воздействие. Но, как я уже говорил, передать посредством такого «мгновенного воздействия» физический сигнал, увы, невозможно.

Согласно принципам теории относительности, физические сигналы (т. е. все, что способно передавать реальную информацию) неизбежно ограничены в своем распространении скоростью света: они могут распространяться медленнее, но быстрее — никогда. Однако для ЭПР-эффектов такое рассмотрение не годится. Представление об ЭПР-эффектах как о конечных сигналах, распространение которых ограничено скоростью света, противоречит всем предсказаниям квантовой теории. (Это обстоятельство хорошо иллюстрируется примером с магическими додекаэдрами — сцепленность между моим додекаэдром и додекаэдром моего коллеги гарантирует их мгновенное взаимодействие, и нет необходимости ждать четыре года, которые затратит на преодоления расстояние между нами световой сигнал; см. §§ 5.3, 5.4, а также примечание 4 в конце главы.) Следовательно, ЭПР-эффекты не могут быть сигналами в обычном смысле этого слова.

Как же в таком случае объяснить тот факт, что ЭПР-эффекты способны-таки повлечь за собой вполне наблюдаемые

последствия? То, что они способны, следует, например, из знаменитой теоремы Джона Белла (см. §5.4). Совместные вероятности, предсказываемые квантовой теорией для различных возможных измерений состояния двух частиц со спином $\frac{1}{2}$ (с независимым выбором направления оси спина левой и правой частицы), невозможно получить ни в какой классической модели несообщающихся левого и правого объектов. (Такого рода примеры описаны и в НРК, с. 284–285 и 301.) Магические додекаэдры из §5.3 дают еще более сильный эффект — здесь речь идет уже не просто о вероятностях, но о вполне точных «да/нет»-ограничениях. Таким образом, хотя левая и правая частицы не *сообщаются* друг с другом в смысле реальной возможности мгновенной передачи сообщений от одного к другому, они, тем не менее, остаются *сцепленными* в том смысле, что их нельзя рассматривать как отдельные независимые объекты, — до того момента, пока их окончательно не расцепит измерение. Квантовая сцепленность — это загадочный феномен, находящийся где-то между прямым сообщением и полным разделением и не имеющий классического аналога. Более того, эффект сцепленности не ослабевает с увеличением расстояния между объектами (в отличие, скажем, от гравитационного или электрического притяжения, величина которого обратно пропорциональна этому самому расстоянию). Эйнштейна это свойство сцепленности крайне нервировало, он называл его «жутковатым действием на расстоянии» (см. [259]).

Квантовая сцепленность не обращает никакого внимания не только на разделенность в пространстве, но и на разделенность во времени. Если измерение одного из компонентов ЭПР-пары выполнено *прежде* такого же измерения другого компонента, то в обычном квантовомеханическом описании считается, как правило, что расцепленность пары явилась результатом именно первого измерения, второе же измерение «захватывает» уже только один, расцепленный, компонент — собственно тот, над которым оно производится. Однако в точности такие же наблюдаемые результаты мы получим, если допустим, что *второе* измерение каким-то образом ретроактивно вызвало расцепление, оставив первое в стороне. Окончательный результат не зависит от порядка выполнения измерений — иначе говоря, измерения *коммутируют* (см. §5.14).



Такая симметрия является необходимым свойством ЭПР-измерений — в противном случае, они противоречили бы наблюдаемым результатам специальной теории относительности. Измерения, производимые над пространственноподобно разделенными событиями (например, событиями, находящимися вне световых конусов друг друга; см. рис. 5.25 и объяснение, приведенное в §4.4), *неминуемо* должны коммутировать — при этом и в самом деле абсолютно неважно, какое именно измерение мы будем полагать «первым», — согласно незыблемым принципам специальной теории относительности. Для того, чтобы в этом убедиться, предположим, что вся эта физическая ситуация описывается с точек зрения двух разных наблюдателей, движущихся каждый в своей системе отсчета (см. рис. 5.26, а также НРК, с. 287). (Эти «наблюдатели» вовсе не обязаны иметь какое бы то ни было отношение к тем, кто выполняет измерения.) В представленной ситуации наблюдатели получат совершенно противоположные представления о том, какое измерение было в действительности выполнено «первым». В отношении измерений ЭПР-типа, феномен квантовой сцепленности — или, если угодно, *рас*цепленности[10] — не знает ни разделенности в пространстве, ни последовательности во времени!

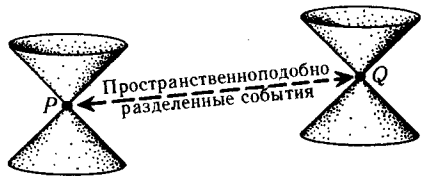



Рис. 5.25. Два события в пространстве-времени называются пространственноподобно разделенными, если каждое из них находится вне светового конуса другого (см. также рис. 4.1, с. 349). В этом случае события не могут оказывать друг на друга никакого причинно-следственного воздействия, следовательно, измерения, производимые над этими событиями, должны коммутировать.

[10] Можно привести примеры [393], когда сцепленность пары частиц сама может оказаться компонентом сцепленной пары!

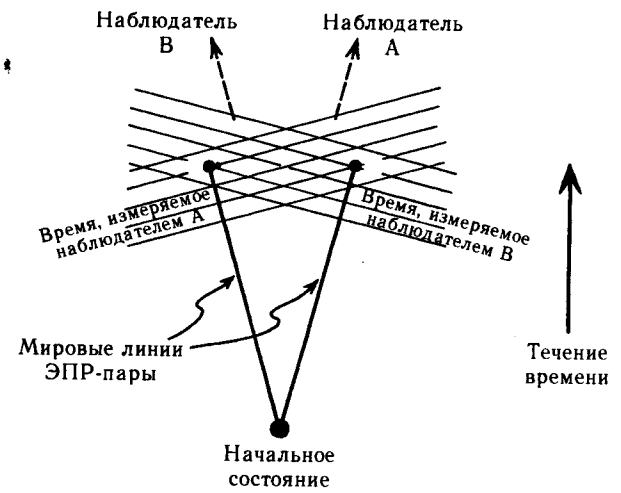



Рис. 5.26. Согласно специальной теории относительности, наблюдатели A и B, движущиеся относительно друг друга, получают различные представления о том, какое из двух пространственноподобно разделенных событий P и Q произошло первым (наблюдатель A полагает, что первым было событие Q, а наблюдатель B уверен, что событие P).

## 5.18. Объяснение загадки магических додекаэдров

Для ЭПР-пары частиц со спином $\frac{1}{2}$ эта пространственная или временна́я нелокальность проявляется исключительно в виде *вероятностей*. Однако на деле феномен квантовой сцепленности вероятностями не ограничивается — он гораздо более конкретен и точен. Магические додекаэдры (и кое-какие более ранние конфигурации[10]) убедительно показывают, что странная нелокальность квантовой сцепленности *не только* порождает вероятности, но и является причиной вполне определенных «да/нет»-эффектов, которые никакими классическими построениями объяснить невозможно.



Попытаемся разобраться в квантовой механике феномена магических додекаэдров из §5.3. Вспомним, что «Квинтэссенциальные Товары», там, у себя, на Бетельгейзе, взяли систему с общим спином 0 (начальное состояние $|\Omega\rangle$), разделили ее на два атома (каждый со спином $\frac{3}{2}$) и подвесили аккуратно каждый атом в центр додекаэдра. Додекаэдры затем тщательно упаковали и отправили почтой (один — мне, а другой — моему коллеге в систему альфы Центавра), обеспечив при этом полную неизменность спиновых состояний этих самых атомов до тех пор, пока кто-то из нас не выполнит, наконец, измерение спина, нажав на одну из кнопок, размещенных в вершинах додекаэдров. Дело в том, что нажатие на кнопку активирует (скажем, с помощью неоднородного магнитного поля, упомянутого в §5.10) измерение (типа измерения Штерна–Герлаха) атома, расположенного в центре соответствующего додекаэдра, — а возможных результатов измерения частицы со спином $\frac{3}{2}$, как нам известно, всего четыре, и они соответствуют (в случае, если измерительное устройство сориентировано вертикально) четырем взаимно ортогональным состояниям: $|\uparrow\uparrow\uparrow\rangle$, $|\downarrow\uparrow\uparrow\rangle$, $|\downarrow\downarrow\uparrow\rangle$ и $|\downarrow\downarrow\downarrow\rangle$. Различаются эти состояния по местоположению атома после прохождения через устройство в одном из четырех возможных лучей. Однако «Квинтэссенциальные Товары» устроили все таким образом, что при нажатии на любую кнопку измерительное устройство непременно оказывается сориентировано в направлении (от центра додекаэдра) на эту самую кнопку. Звонок звенит (результат **ДА**), если атом при измерении обнаруживается во *втором* из четырех возможных местоположений (см. рис. 5.27). Иначе говоря, ответ **ДА** (для случая, когда устройство ориентировано вертикально) вызывается состоянием $|\downarrow\uparrow\uparrow\rangle$ — звенит звонок, за которым следует впечатляющий фейерверк, — остальные три состояния *никакой* реакции *не* вызывают (ответ **НЕТ**). В случае ответа **НЕТ** три оставшиеся луча сводятся вместе (скажем, посредством изменения направленности неоднородного магнитного поля на обратную), что не сопровождается никакими разрушительными эффектами, — и мы снова можем нажимать на какую-нибудь другую кнопку, выбирая тем самым новое направление изменения поля. Отметим тот факт, что каждое нажатие кнопки является, по сути своей, ***примитивным*** измерением, согласно определению этого термина, данному в §5.13.

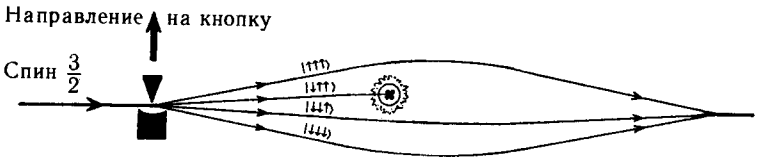



Рис. 5.27. «Квинтэссенциальные Товары» устроили все таким образом, что при нажатии на кнопку в одной из вершин додекаэдра выполняется измерение спина атома со спином $\frac{3}{2}$ в направлении на кнопку (каковое направление принимается за направление «вверх»). Если при этом измерении обнаруживается состояние $|\downarrow\uparrow\uparrow\rangle$, то звенит звонок (результат **ДА**). Если получен результат **НЕТ**, лучи сводятся вместе, и измерение повторяется в каком-либо другом направлении.

Общее состояние $|\Omega\rangle$ нашей системы из двух атомов со спином $\frac{3}{2}$ можно записать следующим образом:

$$|\Omega\rangle=|\mathbf{L}\uparrow\uparrow\uparrow\rangle|\mathbf{R}\downarrow\downarrow\downarrow\rangle-|\mathbf{L}\uparrow\uparrow\downarrow\rangle|\mathbf{R}\downarrow\downarrow\uparrow\rangle+|\mathbf{L}\uparrow\downarrow\downarrow\rangle|\mathbf{R}\downarrow\uparrow\uparrow\rangle-|\mathbf{L}\downarrow\downarrow\downarrow\rangle|\mathbf{R}\uparrow\uparrow\uparrow\rangle.$$

Будем считать мой атом правым; в этом случае, если я обнаруживаю, что он действительно находится в состоянии $|\mathbf{R}\ \downarrow\uparrow\uparrow\rangle$, поскольку звонок звенит при моем первом нажатии на верхнюю кнопку, то звонок додекаэдра моего коллеги должен зазвенеть, если тому случится нажать первой кнопку, противоположную моей, — т. е. состояние его атома $|\mathbf{L}\ \uparrow\downarrow\downarrow\rangle$. Более того, если при нажатии первой кнопки мой звонок не зазвенит, то не зазвенит и его звонок при нажатии противоположной кнопки.

Теперь необходимо убедиться, что при таких примитивных «кнопочных» измерениях действительно выполняются гарантируемые «Квинтэссенциальными Товарами» свойства (а) и (б). В Приложении С приведены некоторые математические подробности предложенного Майораной описания спиновых состояний (в частности, для спина $\frac{3}{2}$), вполне достаточные для какого угодно доказательства. Для упрощения рассуждений представим себе, что сфера Римана проходит через все вершины рассматриваемого додекаэдра, т. е. *описывает* додекаэдр. Отметим далее,



что в описании Майораны **ДА**-состояние для нажатия кнопки в некоторой вершине P додекаэдра включает в себя дважды саму точку P, а также точку P*, антиподальную P, — что и в самом деле соответствует состоянию $|\mathbf{R}\downarrow\uparrow\uparrow\rangle$, если точка P находится на северном полюсе додекаэдра. Иначе говоря, это **ДА**-состояние мы можем обозначить через $|P^*PP\rangle$.

Ключевым свойством спина $\frac{3}{2}$ является то, что **ДА**-состояния для примитивных измерений, соответствующих нажатиям на кнопки при двух «следующих соседних» вершинах, *ортогональны*. В чем тут причина? Покажем, что майорановы состояния $|A^*AA\rangle$ и $|C^*CC\rangle$ действительно ортогональны для любых следующих соседних вершин A и C додекаэдра. Как видно из рис. 5.28, следующими соседними являются вершины додекаэдра, совпадающие с *соседними* вершинами *куба*, вписанного в додекаэдр и имеющего с ним общие центр и восемь вершин. Согласно Приложению С (последний абзац, с. 473), состояния $|A^*AA\rangle$ и $|C^*CC\rangle$ ортогональны, если вершины A и C являются соседними вершинами куба, так что свойство можно считать доказанным.

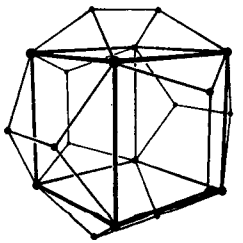


Рис. 5.28. Внутрь правильного додекаэдра можно поместить куб, который будет иметь общие с додекаэдром центр и восемь (из двадцати) вершин. Отметим, что соседние вершины куба являются следующими соседними вершинами додекаэдра.

О чем это нам говорит? В частности, о том, что нажатия кнопок при трех вершинах додекаэдра, соседних с ВЫБРАННОЙ вершиной представляют собой *коммутирующие* измерения (§ 5.14), поскольку по отношению друг к другу эти вершины являются следующими соседними. Таким образом, порядок,

в котором я буду на них нажимать, никак не повлияет на исход дела. Не имеет никакого значения и то, в каком порядке будет нажимать на кнопки своего додекаэдра мой коллега на альфе Центавра. Если его ВЫБРАННОЙ вершиной является вершина, *противоположная* моей, то противоположны моим и три коммутирующие кнопки его додекаэдра. Согласно всему вышесказанному, мой и его звонки должны зазвенеть при нажатии нами на противоположные кнопки независимо от того, в каком порядке каждый из нас нажимает на кнопки своего додекаэдра, — либо ни мой, ни его звонок не зазвенит вообще. Свойство (а) доказано.

Перейдем к свойству (б). Отметим, что гильбертово пространство для спина $\frac{3}{2}$ *четырехмерно*, так что три взаимно ортогональных возможных нажатия, при которых звонок мог бы зазвенеть — скажем, те, которым соответствуют состояния $|A^*AA\rangle$, $|C^*CC\rangle$ и $|G^*GG\rangle$ (в качестве ВЫБРАННОЙ возьмем вершину B), — не вполне исчерпывают всех возможных альтернативных исходов. Остается еще вариант, когда не «звенит» ни одна их этих кнопок, в результате чего мы имеем нулевое измерение (все три кнопки были нажаты, а звонок не прозвенел), т. е. перед нами еще одно состояние (уникальное), ортогональное остальным трем ($|A^*AA\rangle$, $|C^*CC\rangle$, $|G^*GG\rangle$). Обозначим это состояние через $|RST\rangle$, где R, S и T — точки на сфере Римана, необходимые для описания состояния по Майоране. Установить действительное расположение этих трех точек — задача далеко не тривиальная (но вполне решаемая, см. [395]). Впрочем, в настоящий момент нам абсолютно неважно, где именно они располагаются. Достаточно знать, что они где-то на сфере Римана и что их расположение определяется геометрией додекаэдра относительно ВЫБРАННОЙ вершины B. Так, в частности (благодаря симметричности додекаэдра), возьми я в качестве ВЫБРАННОЙ вместо B антиподальную ей вершину B*, тогда результатом отсутствия звонка при нажатии всех кнопок при соседних с B* вершинах A*, C* и G* стало бы состояние $|R^*S^*T^*\rangle$, где R*, S* и T* — точки, антиподальные точкам R, S и T.

Предположим теперь, что мой коллега ВЫБИРАЕТ на своем додекаэдре вершину B, в точности соответствующую той вершине B, что ВЫБРАЛ на своем додекаэдре я. Если при этом его звонок *не* звенит при нажатии любой из трех *его* кнопок при вершинах A, C и G, соседних с B, то *его* измерения (коммутирующие) неизбежно вынуждают *мой* атом перейти в состояние, ортогональное трем состояниям, соответствующим нажатиям на кнопки при *противоположных* вершинах A*, C* и G* моего додекаэдра, т. е. в состояние $|R^*S^*T^*\rangle$. Если же мой звонок также *не звенит*, когда я нажимаю на кнопки при вершинах A, C и G *моего* додекаэдра, то мой атом должен находиться в состоянии $|RST\rangle$. Однако, согласно свойству **C.1** из Приложения C (с. 471), состояние $|RST\rangle$ ортогонально состоянию $|R^*S^*T^*\rangle$; следовательно, невозможно нажать все шесть кнопок без того, чтобы не зазвенел звонок, т. е. свойство (б) также можно считать доказанным.



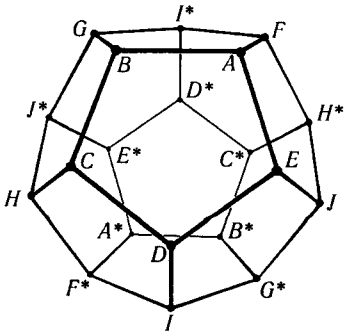



Рис. 5.29. Обозначение вершин додекаэдра, используемое в § 5.18 и Приложении B (с. 467)

Вышесказанное объясняет, каким образом «Квинтэссенциальным Товарам» удается, используя феномен квантовой сцепленности, гарантировать наличие у додекаэдров свойств (а) и (б). Как было показано в § 5.3, *если* бы наши додекаэдры вели себя как *независимые* объекты, из этого немедленно воспоследовали бы «раскрасочные» свойства (в), (г) и (д), что, в свою очередь, привело бы к неразрешимой проблеме раскрашиваемости вершин (каковая неразрешимость явно продемонстрирована в Приложении B, с. 467). Таким образом, то, чего ухитрились добиться с помощью квантовой сцепленности «Квинтэссенциальные Товары», было бы просто-напросто *невозможно*, окажись магические додекаэдры по выходе за ворота фабрики «Квинтэссенциальных Товаров» действительно независимыми объектами, никак

не связанными между собой. Квантовая сцепленность — это не просто досадная морока, не позволяющая нам с легким сердцем игнорировать вероятностные эффекты внешнего окружения на физическую ситуацию. Когда ее влияние удается должным образом обособить, перед нами возникает феномен, точно описываемый математически и зачастую обладающий четкой геометрической организацией.

Предсказания квантовомеханического формализма нельзя описать в терминах объектов, рассматриваемых отдельно один от другого. Феномены квантовой сцепленности невозможно, в общем случае, объяснить рассуждениями «бертлмано-носочного» типа. Следуя правилам стандартной квантовомеханической эволюции — нашей процедуры **U**, — мы приходим к заключению, что «сцепленные» этим диковинным образом объекты остаются сцепленными вне зависимости от того, на какое расстояние им случится удалиться друг от друга. Сцепленность эту может разрушить только процедура **R**. Однако «реальна» ли процедура **R**? Если нет, то сцепленность никуда не исчезает, она остается навечно, пусть и скрытая от наших глаз чрезвычайной сложностью реального мира.

Означает ли это, что всё во Вселенной сцеплено со всем? Как уже было отмечено ранее (см. § 5.17), феномен квантовой сцепленности не похож на феномены, рассматриваемые классической физикой, где интенсивность действия неминуемо убывает на расстоянии, благодаря чему объяснение поведения объектов в лаборатории на Земле не требует от нас знания того, что происходит в данный момент в галактике Туманность Андромеды. Квантовая же сцепленность представляется на первый взгляд как раз тем самым «жутковатым действием на расстоянии», столь раздражавшим Эйнштейна. Однако «действие» это чрезвычайно тонкого рода, и его невозможно использовать для реальной передачи сообщений.

Несмотря на то, что прямого сообщения с ее помощью осуществить не удастся, потенциальные дистанционные («жутковатые») эффекты квантовой сцепленности игнорировать нельзя. Коль скоро сцепленность не разрушается, мы, строго говоря, не можем полагать отдельным и независимым ни один объект во Вселенной. Складывающееся в результате в физической теории положение дел представляется мне весьма далеким от удовлетворительного. Никто не может по-настоящему объяснить, не



выходя за рамки стандартной теории, почему на практике сцепленность *можно*-таки не принимать в расчет. Почему нам вовсе не обязательно представлять Вселенную в виде единого целого, этакого невероятно сложного квантовосцепленного спутанного клубка, не имеющего ничего общего с тем классическим по виду миром, который мы в реальности наблюдаем? На практике квантовые сцепленности разрушаются то и дело применяемой процедурой редукции **R**, что небезуспешно проделали и мы с коллегой, выполнив измерения над сцепленными атомами, помещенными внутрь наших додекаэдров. Является ли, в таком случае, эта самая редукция **R** *реальным* физическим процессом? Иными словами, *действительно* ли **R**, в том или ином смысле, разрушает квантовые сцепления? Или это надо понимать просто как фигуру речи, призванную обозначить некое иллюзорное действие?

В следующей главе мы попытаемся ответить на эти каверзные вопросы. Я убежден, что именно они являются центральными в нашем поиске места невычислимости в физических процессах.

## Примечания

1. См. [296], [299] и [396].
2. Первый проект конкретного эксперимента такого рода был предложен Клаузером, Хорном и Шимони (см. [54] и [55]).
3. Первые эксперименты, результаты которых указывали на подтверждение предсказания квантовой нелокальности, были проведены Фридманом и Клаузером [125]; несколькими годами позже Аспект, Гранжье и Роже [14] получили существенно более полные и однозначные результаты (см. также [13]).
4. Известно еще одно «классическое» объяснение тех ЭПР-эффектов, что наблюдались Аспектом и прочими экспериментаторами. Объяснение это (так называемый «*коллапс с запаздыванием*») предложил Юэн Сквайрс [356], исходя из допущения, что реальные моменты выполнения измерения детекторами в двух удаленных друг от друга точках может разделять довольно существенный промежуток времени. Это допущение рассматривается в контексте некоей теории — само собой, нетрадиционной, вроде тех, что встретятся нам в §§ 6.9 или 6.12, — где делаются вполне конкретные предсказания относительно вероятного момента времени, в который *реально* выполняется каждое из двух квантовых измерений. Поскольку оба эти момента подвержены влиянию всевозможных случайных факторов, ничто не мешает предположить, что один из детекторов

выполнит измерение существенно раньше, чем другой, — настолько раньше, что этого времени вполне хватит на то, чтобы сигнал от первого детектора, распространяясь со скоростью света, достиг второго детектора и передал ему информацию о результате выполненного измерения.

Согласно такой точке зрения, всякое квантовое измерение сопровождается «информационной волной», распространяющейся со скоростью света в направлении от события измерения. Это представление полностью согласуется с классической теорией относительности (см. § 4.4), однако противоречит, на достаточно больших расстояниях, квантовой теории. В частности, коллапсом с запаздыванием невозможно объяснить описанные в § 5.3 свойства магических додекаэдров. Разумеется, соответствующего «эксперимента» пока еще никто не проводил, и можно вполне безнаказанно уверять себя в том, что уж в этом-то случае предсказания квантовой теории нипочем не подтвердятся. У меня, однако, имеется и более серьезное возражение: попытка применения теории «коллапса с запаздыванием» к другим квантовым измерениям сталкивается с серьезными трудностями, приводящими в конечном итоге к нарушению всех стандартных законов сохранения. Например, два достаточно разнесенных детектора смогут при таком раскладе уловить *одну и ту же*, скажем, $\alpha$-частицу, испускаемую при распаде радиоактивного атома, что разом нарушает законы сохранения энергии, электрического заряда и барионного числа! (При достаточно большом расстоянии между детекторами «информационной волне» от первого детектора просто-напросто не хватит времени для того, чтобы успеть «предупредить» второй детектор, запретив ему тем самым принимать ту же $\alpha$-частицу.) Впрочем, «статистически» законы сохранения в данном случае все равно действуют, и мне не известно ни об одном реальном измерении, опровергающем это допущение. Одну из последних оценок статуса соответствующей теории можно найти в [204].

5. Как сообщил мне Абнер Шимони, Кохен и Спекер к тому времени уже самостоятельно пришли к соответствующей переформулировке.
6. Примеры с другими геометрическими конфигурациями можно найти в [305], [260] и [299].
7. Для того чтобы получить самое эффективное «полусеребрёное зеркало», никакого серебра не требуется вовсе, достаточно взять пластину любого прозрачного материала соответствующей толщины, определяемой длиной волны падающего света. Нужный эффект будет достигнут посредством сложной комбинации многократных внутренних отражений и пропусканий, окончательным результатом



чего станут два равных по интенсивности луча света — отраженный и прошедший сквозь. Фазовый сдвиг на четверть длины волны (обусловливающий появление того самого коэффициента $i$) возникает вследствие «унитарности» окончательного разделения исходного луча света на прошедший и отраженный лучи. Более подробное обсуждение имеется в [224].

8. См., например, [94] или [70].
9. Фазовый коэффициент для отраженного состояния я выбрал здесь, в некотором смысле, произвольно. Он частично зависит от того, какого рода зеркало используется. В данном случае, кстати, зеркала могут быть и в самом деле серебрёными, в отличие от «полусеребрёного зеркала» (прекрасно обходящегося вовсе без серебра) в Примечании 7. Выбранный мною коэффициент $i$ представляет собой своего рода компромисс с целью достижения внешнего согласия с коэффициентом, получаемым для «полусеребрёных зеркал». Вообще говоря, до тех пор пока мы остаемся последовательными в отношении *обоих* типов участвующих в эксперименте зеркал, не так уж и важно, какой именно коэффициент выбирается для описания отражения от зеркал непрозрачных.
10. См., например, [225], а также ссылки, перечисленные в примечании 6.

## Приложение В: Нераскрашиваемость додекаэдра

Напомним условие задачи, поставленной в § 5.3. Предлагается показать, что невозможно раскрасить все вершины додекаэдра в БЕЛЫЙ и ЧЕРНЫЙ цвета, соблюдая следующие условия: две «следующие соседние» вершины не могут обе быть БЕЛЫМИ, а шесть вершин, соседних с двумя противоположными (антиподальными) вершинами, не могут быть все ЧЕРНЫМИ. При исключении возможных вариантов раскраски чрезвычайно полезной оказывается симметричность додекаэдра.

Обозначим вершины, как указано на рис. 5.29. Вершины A, B, C, D и E образуют ближайшую к нам пятиугольную грань додекаэдра; дальше, в том же порядке, следуют соседние с ними вершины F, G, H, I и J. Как и в § 5.18, соответствующие антиподальные вершины обозначены через A*, ..., J*. Для начала отметим, что, согласно второму свойству условия, среди вершин додекаэдра хотя бы одна должна быть БЕЛОЙ — пусть это будет A.

Предположим теперь, что среди непосредственных соседей БЕЛОЙ вершины A имеется еще одна БЕЛАЯ вершина — скажем, B (см. рис. 5.29). Тогда все десять вершин, окружающие эту пару, — C, D, E, J, H*, F, I*, G, J* и H — должны быть ЧЕРНЫМИ, так как каждая из них является следующей соседней по отношению либо к A, либо к B. Далее, возьмем шесть вершин, соседних с вершинами из антиподальной пары H, H*. В этой шестерке должна быть хотя бы одна БЕЛАЯ вершина, значит, БЕЛОЙ будет либо F*, либо C* (или обе сразу). Проделав ту же процедуру с парой J, J*, приходим к выводу, что здесь БЕЛОЙ должна быть либо вершина G*, либо E* (или, опять же, обе сразу). Но это *невозможно*! И G*, и E* являются следующими соседними по отношению как к F*, так и к C*. Следовательно, вариант, когда у БЕЛОЙ вершины A имеется БЕЛЫЙ же непосредственный сосед, исключается — в самом деле, ввиду симметричности додекаэдра, невозможной оказывается любая пара соседних БЕЛЫХ вершин.

Таким образом, вершина A должна быть окружена исключительно ЧЕРНЫМИ вершинами B, C, D, E, J, H*, F, I* и G, поскольку каждая из этих вершин является по отношению к A либо соседней, либо следующей соседней. Обратим наше внимание на шесть вершин, соседних с вершинами из антиподальной пары A, A*. Очевидно, что одна из вершин B*, E* или F* должна быть БЕЛОЙ, причем, в силу симметричности додекаэдра, неважно, какая именно, — пусть будет F*. Отметим, что вершины E* и G* являются следующими соседними по отношению к F*, значит, они обе должны быть ЧЕРНЫМИ; ЧЕРНОЙ должна быть и вершина H, поскольку она соседствует с F*, а мы только что исключили возможность существования соседних БЕЛЫХ вершин. Однако так раскрашивать вершины нельзя, потому что при этом *все* соседи антиподальных вершин J, J* оказываются ЧЕРНЫМИ. Вот, собственно, и все доказательство — в *классическом* мире магические додекаэдры невозможны!

# Приложение C: Ортогональность общих спиновых состояний

Предложенное Майораной обобщенное описание спиновых состояний не пользуется широкой известностью среди физиков,



хотя оно весьма удобно и геометрически наглядно. Я расскажу здесь вкратце об основных формулах и о некоторых их геометрических приложениях. Мы, в частности, получим необходимые для рассуждения в §5.18 отношения ортогональности, определяющие геометрию магических додекаэдров. Мои описания существенно отличаются от тех, что первоначально сформулировал Майорана [252], приближаясь, скорее, к описаниям, данным в [299] и [396].

Идея заключается в том, что берется неупорядоченное множество из $n$ точек на сфере Римана, каковые точки рассматриваются как корни комплексного полинома степени $n$, коэффициенты которого, в свою очередь, используются в качестве координат $(n+1)$-мерного гильбертова пространства спиновых состояний (массивной) частицы со спином $\frac{1}{2}n$. Как и в §5.10, основными состояниями будем считать различные возможные результаты измерения спина в вертикальном направлении; представим эти состояния в виде одночленов (добавив соответствующие нормирующие множители, чтобы сохранить единичную длину векторов состояний):

$$
\begin{aligned}
&|\uparrow\uparrow\uparrow\uparrow\ldots\uparrow\uparrow\rangle && — && x^n\\
&|\downarrow\uparrow\uparrow\uparrow\ldots\uparrow\uparrow\rangle && — && n^{1/2}x^{n-1}\\
&|\downarrow\downarrow\uparrow\uparrow\ldots\uparrow\uparrow\rangle && — && \{n(n-1)/2!\}^{1/2}x^{n-2}\\
&|\downarrow\downarrow\downarrow\uparrow\ldots\uparrow\uparrow\rangle && — && \{n(n-1)(n-2)/3!\}^{1/2}x^{n-3}\\
& && \ldots\\
&|\downarrow\downarrow\downarrow\downarrow\ldots\downarrow\uparrow\rangle && — && n^{1/2}x\\
&|\downarrow\downarrow\downarrow\downarrow\ldots\downarrow\downarrow\rangle && — && 1.
\end{aligned}
$$

(Выражения в фигурных скобках — биномиальные коэффициенты.) Таким образом, общее состояние спина $\frac{1}{2}n$,

$$z_0|\uparrow\uparrow\uparrow\ldots\uparrow\rangle+z_1|\downarrow\uparrow\uparrow\ldots\uparrow\rangle+z_2|\downarrow\downarrow\uparrow\ldots\uparrow\rangle+z_3|\downarrow\downarrow\downarrow\ldots\uparrow\rangle+\ldots+ +z_n|\downarrow\downarrow\downarrow\ldots\downarrow\rangle,$$

представляется в виде полинома

$$p(x)=a_0+a_1x+a_2x^2+a_3x^3+\ldots+a_nx^n,$$

где

$$a_0 = z_0,\ a_1 = n^{1/2}z_1,\ a_2 = \{n(n-1)/2!\}^{1/2}z_2,\ \ldots\ a_n = z_n.$$

Корням $x = \alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n$ полинома $p(x) = 0$ соответствуют $n$ точек на сфере Римана, определяющие описание Майораны. Допускается и майоранова точка, задаваемая корнем $x = \infty$, — южный полюс сферы, — это происходит, когда степень полинома $P(x)$ оказывается меньше $n$ на величину, определяемую кратностью этой точки.

Вращение сферы осуществляется посредством следующего преобразования: сначала выполняем замену

$$x \mapsto (\lambda x - \mu)(\overline{\mu}x + \overline{\lambda})^{-1}$$

(где $\lambda\overline{\lambda} + \mu\overline{\mu} = 1$), а затем избавляемся от знаменателей, умножив все выражение на $(\overline{\mu}x + \overline{\lambda})^n$. Таким образом, можно получить полиномы, соответствующие результатам измерений (скажем, с помощью установки Штерна – Герлаха) спина в произвольно выбранном направлении, что дает выражения вида

$$c(\lambda x - \mu)^p(\overline{\mu}x + \overline{\lambda})^{n-p}.$$

Точки, задаваемые отношениями $\mu/\lambda$ и $-\overline{\mu}/\overline{\lambda}$, являются антиподальными на сфере Римана и соответствуют направлению измерения спина и направлению, противоположному ему. (Это предполагает некий подходящий выбор фаз для состояний $|\uparrow\uparrow\uparrow\ldots\uparrow\rangle$, $|\downarrow\uparrow\uparrow\ldots\uparrow\rangle$, $|\downarrow\downarrow\uparrow\ldots\uparrow\rangle$, …, $|\downarrow\downarrow\downarrow\ldots\downarrow\rangle$. Вышеупомянутые свойства и их детальные обоснования удобнее всего рассматривать в терминах 2-спинорного формализма. За подробностями отсылаю читателя к [301], с. 162 и §4.15. Общее состояние спина $\frac{1}{2}n$ описывается там через симметрический $n$-валентный спинор, при этом майораново описание выводится из канонического разложения спинора на симметризованное произведение спиновых векторов.)

Для любой точки $\alpha$ на сфере Римана антиподальной является точка $-1/\overline{\alpha}$. Таким образом, если отразить все майорановы точки, являющиеся корнями полинома

$$a(x) \equiv a_0 + a_1x + a_2x^2 + \ldots + a_{n-1}x^{n-1} + a_nx^n,$$



относительно центра сферы, то мы получим корни полинома

$$a^*(x) \equiv \overline{a}_n - \overline{a}_{n-1}x + \overline{a}_{n-2}x^2 - \ldots - (-1)^n\overline{a}_1x^{n-1} + (-1)^n\overline{a}_0x^n.$$

Пусть состояния $|\alpha\rangle$ и $|\beta\rangle$ заданы, соответственно, полиномами $a(x)$ и $b(x)$, где

$$b(x) \equiv b_0 + b_1x + b_2x^2 + \ldots + b_{n-1}x^{n-1} + b_nx^n;$$

тогда их скалярное произведение имеет вид

$$\langle\beta|\alpha\rangle = \overline{b}_0a_0 + \frac{1}{n}\overline{b}_1a_1 + \frac{2!}{n(n-1)}\overline{b}_2a_2 + \frac{3!}{n(n-1)(n-2)}\overline{b}_3a_3 + \ldots + \overline{b}_na_n.$$

Это выражение инвариантно относительно вращений сферы, что можно непосредственно доказать, используя вышеприведенные формулы.

Применим полученное выражение для скалярного произведения к конкретному случаю $b(x) = a^*(x)$, т. е. к случаю двух состояний, майораново описание одного из которых состоит исключительно из точек, антиподальных точкам, составляющим майораново описание другого. Их скалярное произведение равно (с точностью до знака)

$$a_0a_n - \frac{1}{n}a_1a_{n-1} + \frac{2!}{n(n-1)}a_2a_{n-2} - \ldots - (-1)^n\frac{1}{n}a_{n-1}a_1 + (-1)^na_na_0.$$

Нетрудно заметить, что при *отрицательном* $n$ все члены выражения взаимно уничтожаются, а значит, можно сформулировать следующую теорему (напомним, что состояние, майораново описание которого имеет вид, скажем, P, Q, …, S, обозначается через $|PQ\ldots S\rangle$; точка, антиподальная X, обозначается X*):

**C.1** Если $n$ нечетно, то состояние $|PQR\ldots T\rangle$ ортогонально состоянию $|P^*Q^*R^*\ldots T^*\rangle$.

Из общего выражения для скалярного произведения можно вывести еще два свойства:

**C.2** Состояние $|PPP\ldots P\rangle$ ортогонально любому из состояний $|P^*AB\ldots D\rangle$.

**C.3** Состояние $|\mathrm{QPP}\ldots\mathrm{P}\rangle$ ортогонально состоянию $|\mathrm{ABC}\ldots\mathrm{E}\rangle$ в тех случаях, когда стереографическая проекция (из P*) точки Q* совпадает с центром тяжести множества стереографических проекций (из P*) точек A, B, C, ..., E.

(Центром тяжести множества точек называют центр тяжести совокупности равных точечных масс, размещенных в этих точках. О стереографических проекциях мы говорили в § 5.10, рис. 5.19.) Для доказательства **C.3** развернем сферу так, чтобы точка P* стала ее южным полюсом. Тогда состоянию $|\mathrm{QPP}\ldots\mathrm{P}\rangle$ соответствует полином $x^{n-1}(x \doteq \chi)$, где $\chi$ определяет точку Q на сфере Римана. Вычислив скалярное произведение этого состояния с состоянием, представленным полиномом $(x-\alpha_1)(x-\alpha_2)(x-\alpha_3)\ldots(x-\alpha_n)$, майораново описание которого составляют корни $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n$, находим, что это произведение обращается в нуль, когда

$$1+n^{-1}\overline{\chi}(\alpha_1+\alpha_2+\alpha_3+\ldots+\alpha_n)=0,$$

т. е. когда $-1/\overline{\chi}$ равно $(\alpha_1+\alpha_2+\alpha_3+\ldots+\alpha_n)/n$, иначе говоря, когда точка $-1/\overline{\chi}$ является центром тяжести (на комплексной плоскости) множества точек $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n$. Что и доказывает свойство **C.3**. Для того чтобы доказать **C.2**, поместим в южный полюс точку P. Тогда состоянию $|\mathrm{PPP}\ldots\mathrm{P}\rangle$ соответствует постоянная величина, 1. Если рассматривать ее как полином степени $n$, то соответствующее скалярное произведение обращается в нуль, когда

$$\alpha_1\alpha_2\alpha_3\ldots\alpha_n=0,$$

т. е. когда хотя бы одна точка из множества $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n$ равна 0 или, что то же самое, совпадает с северным полюсом сферы — в данном случае, с точкой P*. Что, собственно, и требовалось доказать.

Свойство **C.2** позволяет интерпретировать майорановы точки в физическом смысле. Исходя из него, можно предположить, что эти точки определяют направления, измерение (типа измерения Штерна–Герлаха) спина в которых дает нулевую вероятность того, что полученное в результате измерения направление оси спина окажется диаметрально противоположным тому направлению, в котором это измерение выполнялось (см. НРК, с. 273). Кроме того, из **C.2** можно вывести свойство для частного



случая: если спин равен $\frac{1}{2}$ $(n=1)$, то ортогональными являются исключительно те состояния, майорановы точки которых антиподальны. Свойство **C.3** позволяет получить общую геометрическую интерпретацию ортогональности в случае спина 1 $(n=2)$. Примечателен частный случай, когда имеются два состояния, представленные в виде двух пар антиподальных точек, причем прямые, соединяющие эти точки, пересекаются в центре сферы под прямым углом. В случае спина $\frac{3}{2}$ $(n=3)$ свойства **C.3** (с некоторой оглядкой на **C.1**) вполне достаточно для подкрепления объяснений, предложенных в § 5.18. (Геометрическую интерпретацию ортогональности в общем случае я здесь давать не буду; может быть, как-нибудь в другой раз.)

Упоминаемое в § 5.18 частное следствие из **C.3** относится к частному случаю, когда P и Q являются соседними вершинами куба, вписанного в сферу Римана, т. е. PQ и Q*P* — противоположные ребра этого куба. Длина отрезка PQ* (или QP*) равна длине PQ (или P*Q*), умноженной на $\sqrt{2}$. Посредством несложных геометрических рассуждений можно показать, что состояния $|\mathrm{P^*PP}\rangle$ и $|\mathrm{Q^*QQ}\rangle$ ортогональны.

# 6

# КВАНТОВАЯ ТЕОРИЯ И РЕАЛЬНОСТЬ

## 6.1. Является ли R реальным процессом?

В предыдущей главе мы сделали попытку понять и принять головоломные **Z**-загадки квантовой теории. Не все эти феномены получили на настоящий момент экспериментальное подтверждение — например, квантовая сцепленность на расстоянии нескольких световых лет[1], — и тем не менее, уже накопленных экспериментальных данных, свидетельствующих о существовании такого рода эффектов, вполне достаточно, чтобы убедиться в том, что **Z**-загадки и в самом деле следует принимать всерьез, поскольку они отражают истинные аспекты поведения самых разных объектов, составляющих тот мир, в котором мы живем.

Процессы, протекающие в нашем физическом мире на квантовом уровне, действительно не поддаются интуитивному осмыслению и во многом совершенно отличны от «классического» поведения, которое мы наблюдаем на более привычном уровне восприятия. Эффекты квантовой сцепленности на расстоянии нескольких метров являются неотъемлемой частью квантового поведения окружающих нас объектов — по крайней мере, это справедливо для объектов квантового уровня (таких, как электроны, фотоны, атомы и молекулы). Контраст между этим странным *квантовым* поведением «микроскопических» объектов (пусть и разделенных вполне макроскопическим расстоянием) и более привычным *классическим* поведением объектов «больших» лежит в основе проблемы **X**-загадок квантовой теории. Может ли, в самом деле, *один* физический закон выступать



в двух различных ипостасях — каждая для «своего» уровня феноменов?

Такое предположение несколько расходится с тем, что мы обычно ожидаем от физического закона. Одним из величайших достижений физики семнадцатого века стала динамика Галилея – Ньютона, согласно которой движение небесных тел подчиняется в точности тем же законам, что управляют движением объектов у нас дома, на Земле. Со времен древних греков (или еще более ранних) ученые полагали, что в небе должны действовать одни законы, а на Земле — другие. Галилей же с Ньютоном смогли показать, что законы одни и те же, различия исключительно в масштабе — фундаментальное прозрение, роль которого в развитии науки переоценить невозможно. Тем не менее (как указывает профессор Иэн Персивал из Лондонского университета), в отношении квантовой теории мы, похоже, решили перенять образ мышления древних греков — один набор законов у нас работает на классическом уровне, а другим, совершенно на первый непохожим, мы пользуемся для описания процессов на квантовом уровне. Я придерживаюсь мнения — и это мнение разделяет, если можно так выразиться, весьма представительное меньшинство физиков, — что такое состояние научной мысли является не чем иным, как временным ступором, и можно предположить, что отыскание соответствующих квантово-классических законов, действующих единообразно на *всех* уровнях феноменов, возвестит научный прорыв, сравнимый по масштабу с тем, у истоков которого стояли Галилей и Ньютон.

Читатель, впрочем, может вполне резонно поинтересоваться, действительно ли та картина, которую дает стандартная квантовая теория для феноменов квантового уровня, не годится для объяснения и классических феноменов. Я убежден, что нет; однако многие склонны это мое убеждение оспаривать, утверждая, что поведение больших или сложных (в некотором смысле) физических систем, каждый из компонентов которых действует в полном согласии с законами квантового уровня, в сущности совпадает с поведением классических объектов (если и не абсолютно, то с очень высокой степенью точности). Попробуем для начала выяснить, можно ли счесть это утверждение — суть которого заключается в том, что наблюдаемое «классическое» поведение макроскопических объектов есть следствие совокупного квантового поведения их микроскопических составляющих, — хоть сколько-

нибудь правдоподобным. Если обнаружится, что нельзя, то нам придется поискать другой путь, который, быть может, приведет нас к более последовательному выводу, имеющему смысл на *всех* уровнях феноменов. Мне, впрочем, следует предупредить читателя о том, что вся эта тема буквально кишит противоречиями. Существует множество самых разнообразных точек зрения, и пытаться дать всесторонний обзор их всех было бы с моей стороны крайне неблагоразумно, не говоря уже о том, чтобы представить детальное опровержение тех из них, что я нахожу невероятными или несостоятельными. Я прошу читателя отнестись снисходительно к тому, что точки зрения, о которых я таки упомяну, будут во многом изложены так, как они выглядят с моей собственной колокольни. Очевидно, что я не смогу сохранить полную беспристрастность, говоря о людях, мнение которых настолько чуждо моему, поэтому я хочу заранее попросить прощения за все те, возможно несправедливые, слова, которые я скажу.

Первая фундаментальная трудность связана с отысканием четкой границы, где *квантовые* процессы, характеризующиеся сохранением суперпозиций различных альтернативных возможностей, действительно переходят — под действием редукции **R** — в процессы *классического* уровня, на котором суперпозиции, по-видимому, невозможны. Трудность эта является результатом свойственной процедуре **R** «скользкости» (с точки зрения наблюдателя), которая не дает нам обнаружить, когда именно она «происходит» — из-за этого, в частности, многие физики вообще не считают редукцию реальным феноменом. Судя по имеющимся данным, результат эксперимента никак не зависит от того, на каком уровне выполняется процедура **R** — необходимо лишь, чтобы этот уровень был выше, чем тот, на котором наблюдались эффекты квантовой интерференции, но ниже, чем тот, на котором мы можем непосредственно воспринимать вместо комплексных линейных суперпозиций реализовавшиеся благодаря редукции классические альтернативы (хотя, как мы вскоре увидим, некоторые физики полагают, что и на этом этапе суперпозиции сохраняются).

Как можно установить, на каком уровне *действительно* происходит редукция — если она, конечно, вообще происходит в физическом смысле? Какой физический эксперимент необходимо поставить для того, чтобы отыскать ответ на этот вопрос? Если **R** — *физический процесс*, то он может происходить на



любом уровне из огромного множества возможных между микроскопическими уровнями *наблюдаемой* квантовой интерференции и макроскопическими уровнями классического *непосредственного* восприятия. Более того, эти различия в «уровнях», похоже, не связаны напрямую с физическими размерами — квантовая сцепленность, например (см. § 5.4), с легкостью «растягивается» до нескольких метров. Мы вскоре покажем, что более подходящей, нежели физические размеры, мерой является в данном случае, *разность энергий*. Как бы то ни было, на нашей, «макроскопической», стороне процесса то место, где «остановится шарик», определяется исключительно нашим же *сознательным восприятием*. С точки зрения физической теории это весьма неудобно, так как нам доподлинно не известно, какие именно физические процессы в мозге отвечают за восприятие. Тем не менее, сама физическая природа этих процессов, похоже, дает для любой теории *реальной* редукции **R** макроскопический предел. Впрочем, и здесь диапазон возможных вариантов между двумя крайностями чрезвычайно велик, что способствует формированию самых разнообразных позиций в отношении того, что же *на самом деле* происходит в тот момент, когда на сцену выходит процедура **R**.

Одним из важнейших является вопрос о «реальности» квантового формализма — или даже квантового мира вообще. Не могу удержаться и не процитировать в этой связи одно замечание профессора Чикагского университета Боба Уолда. Несколько лет назад на одном из банкетов он сказал мне:

> «Если вы и вправду верите в квантовую механику, значит, всерьез вы ее не принимаете».

Мне кажется, что в этом замечании содержится некая глубокая истина как о самой квантовой теории, так и об отношении к ней людей. Те из адептов теории, кто особенно яростно отрицает необходимость какой бы то ни было ее модификации, *не* склонны полагать, что она описывает действительное поведение «реального» квантового мира. Нильс Бор, один из создателей и выдающийся интерпретатор квантовой теории, придерживался в этом отношении наиболее непримиримой позиции. Вектор состояния он, судя по всему, считал не более чем удобной условностью, полезной лишь для вычисления вероятностей результатов допускаемых системой «измерений». Сам по себе вектор состояния и

*не* должен давать объективного описания той или иной квантовой *реальности*, он призван лишь олицетворять «наше знание» о системе. В самом деле, разве можно всерьез полагать, будто понятие «реальность» осмысленно применимо к происходящим на квантовом уровне процессам? Бор, несомненно, принадлежал к тем, кто «и вправду верит в квантовую механику», и, на его взгляд, вектор состояния как раз и не следовало «принимать всерьез» в качестве средства описания физической реальности на квантовом уровне.

Общая альтернатива этой квантовомеханической точке зрения заключается в предположении, что вектор состояния дает-таки строгое математическое описание реального квантового мира — мира, эволюционирующего по чрезвычайно точным законам, хотя, возможно, и не в полном соответствии с математическими правилами, задаваемыми уравнениями квантовой теории. Отсюда, как мне представляется, открываются два основных пути. Одни ученые полагают, что процедура **U** исчерпывающе описывает все, что связано с эволюцией квантового состояния. Процедура же **R**, соответственно, рассматривается как своего рода иллюзия, условность или аппроксимация, но *ни в коем случае* не как часть *действительной* эволюции реальности, описываемой квантовым состоянием. Такое мнение, на мой взгляд, ведет в направлении так называемой концепции *множественности миров*, или *интерпретации Эверетта*[(2)]. Об этой концепции мы поподробнее поговорим буквально через минуту. Другие — как раз те, кто принимает квантовый формализм в наибольшей степени «всерьез», — уверены, что *обе* процедуры, как **U**, так и **R**, представляют (с достаточно большой степенью точности) *действительное* физическое поведение *физически реального*, описываемого вектором состояния, квантового/классического мира. Однако если принимать квантовый формализм настолько всерьез, становится очень нелегко искренне верить в то, что существующая квантовая теория целиком и полностью верна на всех уровнях. Взять хотя бы то, что процедура **R**, в ее теперешнем определении, противоречит многим свойствам процедуры **U**, в частности, *линейности* последней. В этом смысле, разумеется, продолжать и далее «вправду верить в квантовую механику» невозможно. В последующих параграфах мы обсудим упомянутые точки зрения более основательно.



## 6.2. О множественности миров

Попробуем для начала выяснить, насколько далеко мы сможем уйти, следуя первым из «реалистических» путей — тому, что ведет в конечном счете к представлению о существовании «множественных» миров. За истинное описание реальности здесь принимается вектор состояния, эволюционирующий исключительно под действием процедуры **U**. Отсюда неизбежно следует, что законам квантовой линейной суперпозиции должны подчиняться и объекты классического уровня (такие, как бильярдные шары или даже люди). Можно предположить, что никаких серьезных проблем в связи с этим возникнуть не должно, поскольку такие суперпозиции состояний на классическом уровне — явление чрезвычайно редкое, и это еще слабо сказано. Проблема, однако, есть и связана она с *линейностью* эволюции **U**. Под действием **U** весовые коэффициенты состояний в суперпозиции всегда остаются *одинаковыми*, вне зависимости от того, какое количество вещества участвует в процессе. Сама по себе процедура **U** не способна, если можно так выразиться, «разделить» суперпозицию состояний только потому, что система выросла в размерах или усложнилась. Суперпозиции при этом отнюдь не проявляют тенденции к «исчезновению» при переходе на классический уровень, в результате чего выраженные суперпозиции состояний классических объектов должны стать не менее распространенным феноменом, нежели суперпозиции квантовых состояний. Отсюда неизбежно следует вопрос: почему в таком случае мы, воспринимая мир классических объектов, не сталкиваемся с такими макроскопическими суперпозициями альтернативных состояний ежедневно?

У приверженцев концепции множественности миров имеется на этот счет объяснение. Попробуем в нем разобраться. Представим себе ситуацию, подобную той, что мы рассматривали в § 5.17, — детектор фотонов, описываемый состоянием $|\Psi\rangle$, оказывается на пути фотона, находящегося в суперпозиции состояний $|\alpha\rangle + |\beta\rangle$, причем $|\alpha\rangle$ активирует детектор, $|\beta\rangle$ же оставляет все как есть. (Возможно, фотон, испущенный некоторым источником, успел по пути встретиться с полупрозрачным зеркалом, и состояния $|\alpha\rangle$ и $|\beta\rangle$ описывают, соответственно, пропущенную и отраженную части общего состояния фотона.) Мы здесь не говорим о применимости концепции вектора состояния к объектам

классического уровня (весь детектор целиком), так как в рамках данной точки зрения векторы состояния являются точными представлениями реальности на всех ее уровнях. Таким образом, состояние $|\Psi\rangle$ может описывать весь детектор целиком, а не только лишь некоторые квантовые его элементы, первыми встречающие фотон, как было в §5.17. Отметим, что, как и в §5.17, после собственно момента столкновения состояния детектора и фотона эволюционируют из произведения $|\Psi\rangle(|\alpha\rangle + |\beta\rangle)$ в сцепленное состояние

$$|\Psi_Д\rangle + |\Psi_Н\rangle|\beta'\rangle.$$

*Реальность* описывается теперь вот этим вот сцепленным состоянием, рассматриваемым как *единое целое*. Мы *не* говорим: «*либо* детектор зарегистрировал и поглотил фотон (состояние $|\Psi_Д\rangle$), либо детектор фотона не зарегистрировал, и фотон остался свободным (состояние $|\Psi_Н\rangle|\beta'\rangle$)». Вместо этого мы говорим: «*обе* альтернативы сосуществуют в суперпозиции, как часть всеобщей реальности, в которой *все* такие суперпозиции сохраняются». Можно распространить ситуацию и вообразить себе экспериментатора-человека, который разглядывает детектор с целью выяснить, зарегистрировал ли тот прибытие фотона. Прежде чем обратить свой взор к детектору, человек также должен был пребывать в некотором квантовом состоянии, скажем, $|\Sigma\rangle$; таким образом, мы получаем на данном этапе следующее совокупное «произведение» состояний:

$$|\Sigma\rangle(|\Psi_Д\rangle + |\Psi_Н\rangle|\beta'\rangle).$$

Далее, изучив состояние детектора, наблюдатель каким-то образом воспринимает, что либо детектор зарегистрировал и поглотил фотон (состояние $|\Sigma_Д\rangle$), либо детектор фотона не зарегистрировал (ортогональное состояние $|\Sigma_Н\rangle$). Если допустить, что наблюдатель не взаимодействует с детектором после наблюдения, то ситуация описывается следующим вектором состояния:

$$|\Sigma_Д\rangle|\Psi'_Д\rangle + |\Sigma_Н\rangle|\Psi'_Н\rangle|\beta''\rangle.$$

То есть теперь у нас имеется два различных (ортогональных) состояния наблюдателя, каждое из которых вносит свой вклад в общее состояние системы. Согласно первому, наблюдатель находится в состоянии восприятия регистрации детектором прибытия фотона; это состояние сопровождается состоянием детектора, при котором фотон действительно регистрируется. Согласно

Наблюдатель никогда не "видит" суперпозиции макросостояний



же второму, наблюдатель находится в состоянии восприятия отсутствия регистрации детектором прибытия фотона; это состояние сопровождается состоянием детектора, при котором фотон не регистрируется, и состоянием фотона, свободно улетающего прочь. При этом, в соответствии с концепцией множественности миров, в рамках одного общего состояния сосуществуют различные экземпляры (варианты, копии) «Я» наблюдателя, располагающие различным опытом восприятия окружающего мира. Действительное состояние мира, окружающего каждый экземпляр, будет соответствовать опыту восприятия, которым этот экземпляр располагает.

Это представление можно обобщить на более «реалистичные» физические ситуации, где одновременно сосуществуют уже не два возможных варианта развития событий, как в приведенном примере, а огромные количества различных квантовых альтернатив, непрерывно возникающих на протяжении всей истории Вселенной. Таким образом, общее состояние Вселенной действительно объединяет в себе множество различных «миров», а любой наблюдатель-человек существует во множестве различных экземпляров сразу. Каждый экземпляр воспринимает тот мир, который не противоречит его собственному опыту восприятия, при этом нас с вами хотят убедить в том, что для построения удовлетворительной теории ничего больше и не нужно. Процедура **R**, согласно такой точке зрения, оказывается *иллюзией*, возникающей как следствие некоторых особенностей восприятия квантовосцепленного мира макроскопическим наблюдателем.

Что касается меня, то должен сказать, что я вообще не нахожу эту точку зрения сколько-нибудь удовлетворительной. И дело здесь не столько в исключительной расточительности такой картины мира — хотя это и само по себе уже достаточно подозрительно, если не сказать больше. Более серьезное возражение состоит в том, что концепция множественности миров не дает *настоящего* решения «проблемы измерения», т. е. не достигает цели, ради которой была создана.

Решить *проблему квантового измерения* — значит понять, каким образом макроскопическое поведение в **U**-эволюционирующих квантовых системах порождает (или *эффективно* порождает) в качестве своего свойства процедуру **R**. Эта проблема не решается простым указанием на возможный сценарий, предположительно допускающий **R**-подобное поведение. Необ-

И мой опыт не содержит суперпозиций макросостояний.

ходима теория, позволяющая хоть как-то понять, какие именно *обстоятельства* вызывают к жизни процедуру **R** (или, на худой конец, ее иллюзию). Более того, необходимо найти объяснение той замечательной *точности*, с которой работает процедура **R**. Судя по всему, люди склонны полагать, что вся точность квантовой теории заключена в ее динамических уравнениях — в эволюции $\mathbf{U}_t$. Однако и редукция **R** сама по себе ничуть не менее точна в предсказании вероятностей, и до тех пор, пока мы не поймем, каким образом ей это удается, удовлетворительной теории у нас не будет.

Поскольку ничего большего концепция множественности миров не предлагает, действительного и исчерпывающего объяснения ни одному из этих феноменов мы не получаем. В отсутствие теории, описывающей, каким образом «воспринимающее сознание» разделяет мир на ортогональные альтернативы, у нас нет никаких причин ожидать, что такое сознание не будет способно осознавать линейные суперпозиции совершенно различных состояний теннисных мячей или, скажем, слонов. (Следует отметить, что одна лишь *ортогональность* «воспринимаемых состояний» — например, состояний $|\Psi_\text{Д}\rangle$ и $|\Psi_\text{Н}\rangle$ в приведенном выше примере — никоим образом не помогает эти состояния разделить. Сравните, например, пару состояний $|\mathbf{L}\leftarrow\rangle$ и $|\mathbf{L}\rightarrow\rangle$ с парой $|\mathbf{L}\uparrow\rangle$ и $|\mathbf{L}\downarrow\rangle$, которыми мы пользовались при обсуждении ЭПР-феноменов в §5.17. Обе пары состояний ортогональны, точно так же как ортогональны состояния $|\Psi_\text{Д}\rangle$ и $|\Psi_\text{Н}\rangle$, однако выбрать одну пару в ущерб другой мы не можем.) И еще одно: концепция множественности миров никак не объясняет чрезвычайную точность того удивительного правила, которое чудесным образом превращает квадраты модулей комплексных весовых коэффициентов в относительные вероятности[3]. (См. также §§6.6 и 6.7.)

## 6.3. Не принимая вектор $|\psi\rangle$ всерьез

Существует много различных вариантов точки зрения, согласно которой вектор состояния $|\psi\rangle$ *не следует* рассматривать как действительное отображение той или иной физической реальности, существующей на квантовом уровне. Вектор $|\psi\rangle$ вводится лишь в качестве вычислительного приема, удобного исключительно для вычисления вероятностей, либо служит для выражения «состояния знания» экспериментатора о физической системе. Иногда под $|\psi\rangle$ понимается не состояние индивидуальной физической системы, но целый *ансамбль* возможных подобных физических систем. Часто утверждают, что поведение вектора сложносцепленного состояния $|\psi\rangle$ ничем, *с практической точки зрения* (*for all practical purposes*[1], или просто FAPP с легкой руки Джона Белла[4]), не отличается от поведения такого ансамбля физических систем — а большего о проблеме измерения физикам знать и не нужно. Иногда можно услышать, что вектор $|\psi\rangle$ не может описывать какую бы то ни было квантовую реальность, так как понятие «реальность» к феноменам квантового уровня неприменимо — оно теряет здесь всякий смысл, поскольку реальным является лишь то, что можно «измерить».



Многие (в том числе и я — а также Эйнштейн и Шрёдингер, так что компания подобралась очень даже неплохая), впрочем, убеждены, что ничуть не больше смысла в ограничении «реальности» лишь объектами, которые мы способны воспринять — например, при помощи измерительных устройств (некоторых из них, по крайней мере), — и лишении «права на реальность» объектов, существующих на более глубоком, более фундаментальном уровне. Я не сомневаюсь, что мир на квантовом уровне выглядит странно и непривычно, но он отнюдь не становится от этого «нереальным». В самом деле, разве могут реальные объекты состоять из нереальных компонентов? Более того, управляющие квантовым миром математические закономерности замечательно точны — ничуть не менее точны, нежели более привычные уравнения, описывающие поведение макроскопических объектов, — несмотря на все те туманные образы, с которыми в нашем сознании ассоциируются «квантовые флуктуации» и «принцип неопределенности».

Однако убежденность в том, что хоть какая-то реальность должна существовать и на квантовом уровне, не избавляет нас от сомнений в возможности точно описать эту самую реальность посредством вектора состояния $|\psi\rangle$. В доказательство «нереальности» $|\psi\rangle$ выдвигаются самые различные аргументы. Во-первых, вектор $|\psi\rangle$, по всей видимости, вынужден время от времени претерпевать этот загадочный нелокальный разрывный «скачок»,

[1] С практической точки зрения (англ.). — *Прим. перев.*

который я обозначаю здесь буквой **R**. Несколько неподобающее поведение для физически приемлемого описания мира, особенно если учесть, что у нас уже имеется изумительно точное и непрерывное уравнение Шрёдингера **U**, согласно которому, как предполагается, и эволюционирует вектор $|\psi\rangle$ (бóльшую часть времени). Однако, как мы успели убедиться, эволюция **U** сама по себе заводит нас в дебри сложностей и неясностей множественно-мировых интерпретаций; если же мы хотим получить картину, сколько-нибудь адекватно описывающую реальную Вселенную, которая, как нам представляется, нас окружает, то нам просто необходима какая-никакая процедура **R**.

Другое нередко выдвигаемое возражение против реальности вектора $|\psi\rangle$ сводится к следующему: чередование **U**, **R**, **U**, **R**, **U**, **R**, ..., представляющее собой, в сущности, типичное описание процесса в квантовой теории, не симметрично во времени (каждое **U**-действие *начинается* с процедуры **R**, но не завершается ею), и существует другое, полностью эквивалентное первому описание, в котором **U**-эволюции обращены во времени (см. НРК, с. 355, 356; рис. 8.1, 8.2). Почему первое описание соответствует «реальности», а второе нет? Есть мнение, что всерьез следует принимать *оба* описания (как прямую, так и обратную эволюцию вектора состояния) — они сосуществуют и дают в совокупности полное описание физической реальности (см. [61], [381] и [2]). Я склонен думать, что предположения эти, скорее всего, не лишены серьезных оснований, однако в настоящий момент мы на них останавливаться не будем. Мы вкратце коснемся их (и некоторых других родственных им) ниже, в § 7.12.

Одно из наиболее частых возражений против принятия вектора $|\psi\rangle$ всерьез в качестве описания реальных процессов состоит в том, что его нельзя непосредственно «измерить» — в том смысле, что не существует экспериментального способа определить вектор состояния (пусть и с точностью до коэффициента пропорциональности), если мы об этом состоянии ничего не знаем. Возьмем для примера атом со спином $\frac{1}{2}$. Вспомним (§ 5.10, рис. 5.19), что каждое возможное состояние спина такого атома характеризуется каким-то конкретным направлением в обычном пространстве. Однако если мы не имеем ни малейшего понятия, что это за направление, определить его мы никак не сможем. Мы можем лишь выбрать какое-либо одно направление и выяснить,



в этом направлении ориентирована ось спина (**ДА**) или же в противоположном (**НЕТ**). Каким бы ни было начальное состояние спина, соответствующее направление в гильбертовом пространстве проецируется либо в **ДА**-пространство, либо в **НЕТ**-пространство; каждый исход реализуется с вполне определенной вероятностью. И тут мы теряем бóльшую часть информации о том, каким было «действительное» начальное состояние спина. Все, что мы можем получить из измерения направления спина (в случае атома со спином $\frac{1}{2}$), укладывается в *один бит* информации (ответ на общий вопрос — **ДА** или **НЕТ**), тогда как возможные состояния направления оси спина образуют континуум, для точного определения которого потребуется бесконечное количество битов информации.

Все это так, и все же противоположную позицию принять ничуть не легче — ту, согласно которой вектор состояния $|\psi\rangle$ оказывается в некотором роде физически «нереальным», являя собой лишь оболочку, содержащую полную сумму «наших знаний» о физической системе. Я бы даже сказал, что принять эту позицию неимоверно трудно, особенно если учесть, что подобная роль «знания» подразумевает немалую долю субъективности. О *чьем*, в конце концов, знании идет здесь речь? Совершенно точно — не о моем. Я очень мало действительно знаю об отдельных векторах состояния, детально описывающих поведение всех до единого окружающих меня объектов. А они, как ни в чем не бывало, продолжают себе свою идеально организованную деятельность, нимало не заботясь ни о том, что именно может стать кому-то «известно» о том или ином векторе состояния, ни о том, кто же станет счастливым обладателем этого драгоценного знания. Разве разные экспериментаторы, располагающие разным знанием о какой-либо физической системе, описывают эту самую систему с помощью различных векторов состояния? Отнюдь; все возникающие здесь различия относятся к тем особенностям каждого конкретного эксперимента, которые не оказывают сколько-нибудь существенного влияния на конечный результат.

Один из наиболее сильных доводов[5] в опровержение этой субъективной точки зрения на реальность $|\psi\rangle$ следует из того факта, что, каким бы ни был вектор состояния $|\psi\rangle$, всегда возможно (по крайней мере, в принципе) осуществить *примитивное измерение* (см. § 5.13), **ДА**-пространство которого пред-

ставляет собой луч в гильбертовом пространстве, определяемый вектором $|\psi\rangle$. Дело в том, что физическое состояние $|\psi\rangle$ (определяемое лучом комплексных кратных $|\psi\rangle$) определено *однозначно*, в силу того, что результат **ДА** для данного состояния является абсолютно *достоверным*. Никакое другое состояние таким свойством не обладает. Для любого другого состояния речь может идти лишь о некоторой вероятности (всегда меньшей, нежели полная уверенность) получения результата **ДА**, не исключающей и возможности того, что будет получен результат **НЕТ**. Таким образом, хотя мы и не можем посредством какого бы то ни было измерения выяснить, что же такое *в действительности* представляет собой вектор $|\psi\rangle$, физическое состояние $|\psi\rangle$ однозначно определяется тем, что должно (согласно соответствующему вектору) являться результатом измерения, которое *могло бы* быть осуществлено над этим состоянием. Здесь мы вновь встречаемся с контрфактуальностью (см.§§ 5.2, 5.3); впрочем, мы уже видели, насколько важную роль в предсказаниях квантовой теории играют контрфактуальные соображения.

Дабы прибавить нашему рассуждению убедительности, вообразим, что квантовая система установлена в некое известное состояние, скажем, $|\phi\rangle$, и что согласно вычислениям, это состояние по прошествии времени $t$ эволюционирует под действием процедуры **U** в другое состояние, скажем, $|\psi\rangle$. Пусть состояние $|\phi\rangle$ представляет, например, состояние «спин вверх» ($|\phi\rangle = |\uparrow\rangle$) атома со спином $\frac{1}{2}$, и предположим, что система оказалась в этом состоянии под действием какого-то предыдущего измерения. Допустим, что наш атом обладает магнитным моментом, направление которого совпадает с направлением оси спина (т. е. представляет собой маленький магнит, ориентированный в направлении оси спина). Направление же оси спина атома, помещенного в магнитное поле, вполне определенным образом прецессирует, что можно точно вычислить и представить как действие процедуры **U**, переводящее спин за время $t$ в новое состояние, скажем, $|\psi\rangle = |\rightarrow\rangle$. Следует ли это вычисленное состояние принимать всерьез как часть физической реальности? Не вижу причин в этом ему отказывать. Поскольку состояние $|\psi\rangle$ никак не может не учитывать *возможность* того, что нам вдруг взбредет в голову измерить его посредством вышеупомянутого примитивного измерения, того самого измерения, **ДА**-пространство которого



состоит исключительно из кратных вектора $|\psi\rangle$. В данном случае таким измерением является измерение спина в направлении $\rightarrow$. На это измерение система должна давать *уверенный* ответ **ДА**, а этого не может гарантировать никакое состояние спина атома, кроме $|\psi\rangle = |\rightarrow\rangle$.

Можно отыскать множество самых разнообразных физических ситуаций, в которых подобное примитивное измерение окажется практически неосуществимым. И все же стандартные правила квантовой теории предполагают, что *в принципе* такие измерения возможны. Если же мы полагаем, что в случае некоторых «достаточно сложных» разновидностей состояний $|\psi\rangle$ примитивные измерения невозможны в принципе, то нам придется пересмотреть самые основы квантовой теории. Может быть, их и впрямь стоит пересмотреть (некоторые конкретные шаги в этом направлении я предложу в § 6.12). Следует, впрочем, понимать, какого рода пересмотр потребуется, если мы и впредь намерены отрицать *объективные* различия между разными квантовыми состояниями или, что одно и то же, объективную реальность вектора состояния $|\psi\rangle$ в некотором строгом физическом смысле (пусть и с точностью до коэффициента пропорциональности).

В качестве «минимального» пересмотра, затрагивающего лишь теорию измерения, часто предлагают ввести так называемые *правила суперселекции*[6], которые и в самом деле эффективно отрицают возможность выполнения определенных типов примитивных измерений системы. Мне не хочется рассматривать здесь эти правила в подробностях, так как ни одно подобное предложение, насколько мне известно, не дошло в своем развитии до той стадии, на которой можно было бы говорить о формировании сколько-нибудь связной общей позиции в отношении проблемы измерения. Подчеркну лишь, что даже минимальный пересмотр подобного рода все равно остается пересмотром — и лишь подтверждает наличие насущной необходимости в пересмотре теории в целом.

В заключение, думаю, следует упомянуть о том, что существует и множество иных подходов к квантовой механике, которые хоть и не противоречат предсказаниям традиционной теории в принципе, но все же дают «картины реальности», так или иначе отличные от той реальности, где вектор состояния $|\psi\rangle$ «принимают всерьез», полагая, что он эту реальность и представляет. Среди них — *пилотно-волновая* теория Луи де Бройля [77] и

Дэвида Бома [33], нелокальная теория, согласно которой существуют объекты, эквивалентные одновременно волновым функциям *и* системам классических частиц, причем *и те, и другие* полагаются в данной теории «реальными». (См. также [34].) Другие точки зрения (вдохновленные Ричардом Фейнманом и его подходом к квантовой теории [118]) оперируют целыми «историями» возможного поведения — согласно этим точкам зрения, истинная картина «физической реальности» несколько отличается от той, которую дает обыкновенный вектор состояния $|\psi\rangle$. Аналогичной общей позиции, которая, впрочем, учитывает еще и возможность, по сути, многократных частичных измерений (в соответствии с анализом, предпринятым в [4]), придерживаются авторы работ [174], [279] и [141]. Было бы неуместно, как мне кажется, углубляться здесь в обсуждение этих разнообразных альтернативных точек зрения (хотя следует все же упомянуть о том, что формализм матриц плотности, который вводится в следующем параграфе, играет в некоторых из этих теоретических построений не последнюю роль — как и в операторном подходе Хаага [179]). Скажу лишь, что, хотя многое в этих процедурах представляет значительный интерес и обладает некоторой вдохновляющей оригинальностью, я все же совершенно не убежден, что с их помощью можно действительно решить проблему измерения. Разумеется, я могу и ошибаться, но это покажет лишь время.

## 6.4. Матрица плотности

Многие физики, полагая себя людьми практичными, вопросами «реальности» вектора $|\psi\rangle$ не интересуются. От $|\psi\rangle$ им нужно лишь одно — возможность вычислять с его помощью вероятности того или иного дальнейшего физического поведения объекта. Часто бывает так, что состояние, выбранное изначально для представления физической ситуации, приобретает под действием эволюции чрезвычайную сложность, а его сцепленности с элементами окружения становятся настолько запутанными, что на практике совершенно невозможно проследить за эффектами квантовой интерференции, отличающими такое состояние от множества других ему подобных. Все уверения в том, что явившийся результатом данной конкретной эволюции вектор состояния сколько-нибудь более реален, нежели прочие, на практике



от него неотличимые, наши «практичные» физики, без сомнения, сочтут абсолютно лишенными смысла. В самом деле, скажут они, любой *отдельный* вектор состояния, пригодный для описания «реальности», всегда можно заменить подходящей *вероятностной комбинацией* векторов состояния. Если применение процедуры **U** к некоему вектору состояния, представляющему начальное состояние системы, дает результат, *с практической точки зрения* (FAPP-подход Белла) неотличимый от того, что был бы получен с помощью такой вот вероятностной комбинации векторов состояния, то вероятностная комбинация достаточно хороша для описания мира и отыскивать **U**-эволюционировавший вектор состояния нужды нет.

Часто утверждают, что с такими же мерками можно подходить и к процедуре **R** — по крайней мере, на практике (все тот же FAPP). Двумя параграфами ниже мы попытаемся найти ответ на вопрос, можно ли в самом деле разрешить кажущийся **U**/**R**-парадокс одними лишь этими методами. Однако прежде я хотел бы рассказать подробнее о процедурах, принятых в стандартных FAPP-подходах к объяснению **R**-процесса (реального или кажущегося).

Ключевым в этих процедурах является математический объект, называемый *матрицей плотности*. Понятие матрицы плотности играет в квантовой теории весьма важную роль, и именно она, а не вектор состояния, лежит в основе большинства стандартных математических описаний процесса измерения. Центральную роль отводит матрице плотности и мой, менее традиционный, подход, особенно в том, что касается ее связи со стандартными FAPP-процедурами. По этой причине нам, к сожалению, придется углубиться в математический формализм квантовой теории несколько далее, нежели было необходимо прежде. Надеюсь, что читателя-неспециалиста такая перспектива не отпугнет. Даже при отсутствии полного понимания, мне думается, любому читателю будет полезно хотя бы бегло просматривать математические рассуждения по мере их появления — несомненно, со временем придет и осмысление. Это стало бы существенным подспорьем для понимания некоторых из дальнейших аргументов и тонкостей, сопровождающих поиски ответа на вопрос, почему нам действительно и насущно необходима усовершенствованная теория квантовой механики.

В отличие от отдельного единичного вектора состояния, мат-

рицу плотности можно рассматривать как представление комбинации вероятностей нескольких возможных *альтернативных* векторов состояния. Говоря о «комбинации вероятностей», мы подразумеваем лишь, что существует некоторая неопределенность в отношении действительного состояния системы, при этом каждому из возможных альтернативных векторов состояния поставлена в соответствие некоторая вероятность — самая обычная классическая вероятность, выраженная самым обычным вещественным числом. Однако матрица плотности вносит в это описание некоторую путаницу (заложенную изначально), поскольку не отличает *классические* вероятности, фигурирующие в вышеупомянутой взвешенной вероятностной комбинации, от вероятностей *квантовомеханических*, возникающих в результате процедуры **R**. Дело в том, что операционными методами различить эти вероятности невозможно, поэтому в операционном же смысле вполне уместным представляется математическое описание (матрица плотности), которое такого различия *не* делает.

Как выглядит это математическое описание? Я не стану углубляться в ненужные здесь подробности, лишь вкратце изложу основные концепции. Идея матрицы плотности, вообще говоря, весьма изящна[2]. Начать с того, что вместо каждого отдельного состояния $|\psi\rangle$ мы используем объект вида

$$|\psi\rangle\langle\psi|.$$

Что означает такая запись? Не прибегая к точному математическому определению, которое для нас сейчас несущественно, можно сказать, что это выражение представляет собой особого рода «произведение» (точнее, вид тензорного произведения, см. § 5.15) вектора состояния $|\psi\rangle$ и «комплексно сопряженного» ему вектора $\langle\psi|$. Вектор состояния $|\psi\rangle$ мы полагаем *нормированным* (т. е. $\langle\psi|\psi\rangle = 1$); тогда выражение $|\psi\rangle\langle\psi|$ однозначно определяется физическим состоянием, представленным вектором $|\psi\rangle$ (поскольку не зависит от изменений фазового множителя $|\psi\rangle \mapsto e^{i\theta}|\psi\rangle$, см. § 5.10). В системе обозначений Дирака исходный вектор $|\psi\rangle$ называется «кет»-вектором, а соответствующий ему $\langle\psi|$ — «бра»-вектором. Бра-вектор $\langle\psi|$ и кет-вектор $|\phi\rangle$ могут образовывать и *скалярное произведение* («bra-ket»[3]):

$$\langle\psi|\phi\rangle,$$

с таким обозначением мы уже встречались в § 5.12. Значением скалярного произведения является самое обычное комплексное число, тогда как тензорное произведение $|\psi\rangle\langle\phi|$ в матрице плотности дает более сложный математический «объект» — элемент некоторого векторного пространства.

Перейти от непонятного «объекта» к обычному комплексному числу позволяет особая математическая операция, называемая *вычислением следа* (или *суммы элементов главной диагонали*) матрицы. Для простого выражения $|\psi\rangle\langle\phi|$ эта операция сводится к простой перестановке членов, дающей в результате скалярное произведение:

$$\text{СЛЕД}\,(|\psi\rangle\langle\phi|) = \langle\phi|\psi\rangle.$$

В случае суммы членов «след» вычисляется линейно: например,

$$\text{СЛЕД}\,(z|\psi\rangle\langle\phi| + w|\alpha\rangle\langle\beta|) = z\langle\phi|\psi\rangle + w\langle\beta|\alpha\rangle.$$

Я не стану в подробностях выводить все математические свойства таких объектов, как $\langle\psi|$ и $|\psi\rangle\langle\phi|$, однако кое о чем упомянуть стоит. Во-первых, произведение $|\psi\rangle\langle\phi|$ подчиняется тем же алгебраическим правилам, что перечислены на с. 446 для произведения $|\psi\rangle|\phi\rangle$ (за исключением последнего, которое к данному случаю неприменимо):

$$\begin{aligned}(z|\psi\rangle)\langle\phi| &= z(|\psi\rangle\langle\phi|) = |\psi\rangle(z\langle\phi|),\\ (|\psi\rangle + |\chi\rangle)\langle\phi| &= |\psi\rangle\langle\phi| + |\chi\rangle\langle\phi|,\\ |\psi\rangle(\langle\phi| + \langle\chi|) &= |\psi\rangle\langle\phi| + |\psi\rangle\langle\chi|.\end{aligned}$$

Следует также отметить, что бра-вектор $\bar{z}\langle\psi|$ является комплексным сопряженным кет-вектора $z|\psi\rangle$ (поскольку число $\bar{z}$ есть комплексное сопряженное комплексного числа $z$, см. с. 412), а сумма $\langle\psi| + \langle\chi|$ — комплексным сопряженным суммы $|\psi\rangle + |\chi\rangle$.

[2] Эта идея была предложена в 1932 году выдающимся венгерско-американским математиком Джоном фон Нейманом. Ему же, главным образом, мы обязаны теорией, опиравшейся на первопроходческие труды Алана Тьюринга и положившей начало развитию электронных компьютеров. Кроме того, фон Нейман стоял у истоков теории игр (см. ссылку в примечании 9 после третьей главы, с. 335) и, что ближе к теме нашего разговора, первым четко определил две квантовые процедуры, которые я обозначил здесь буквами «**U**» и «**R**».

[3] Созвучно английскому *bracket* «скобка». — *Прим. перев.*

Допустим, нам нужно составить матрицу плотности, представляющую некоторую комбинацию вероятностей нормированных состояний, скажем, $|\alpha\rangle$ и $|\beta\rangle$; вероятности, соответственно, равны $a$ и $b$. Правильная матрица плотности в данном случае будет иметь вид

$$\boldsymbol{D} = a|\alpha\rangle\langle\alpha| + b|\beta\rangle\langle\beta|.$$

Для трех нормированных состояний $|\alpha\rangle$, $|\beta\rangle$, $|\gamma\rangle$ с соответствующими вероятностями $a$, $b$, $c$ имеем

$$\boldsymbol{D} = a|\alpha\rangle\langle\alpha| + b|\beta\rangle\langle\beta| + c|\gamma\rangle\langle\gamma|,$$

и так далее. Из того, что вероятности всех альтернативных вариантов должны в сумме давать единицу, можно вывести важное свойство, справедливое для любой матрицы плотности:

$$\text{СЛЕД}\,(\boldsymbol{D}) = 1.$$

Как же использовать матрицу плотности для вычисления вероятностей, результатов измерения? Рассмотрим сначала простой случай примитивного измерения. Спросим, находится ли система в физическом состоянии $|\psi\rangle$ (**ДА**) или в ином состоянии, ортогональном $|\psi\rangle$ (**НЕТ**). Само измерение представляет собой математический объект (так называемый *проектор*), очень похожий на матрицу плотности:

$$\boldsymbol{E} = |\psi\rangle\langle\psi|.$$

Вероятность $p$ получения ответа **ДА** определяется из выражения

$$p = \text{СЛЕД}\,(\boldsymbol{DE}),$$

где произведение $\boldsymbol{DE}$ само представляет собой объект, подобный матрице плотности. Оно вычисляется с помощью несложных алгебраических правил, необходимо лишь соблюдать порядок «умножений». Например, для вышеприведенной двучленной суммы $\boldsymbol{D} = a|\alpha\rangle\langle\alpha| + b|\beta\rangle\langle\beta|$ имеем

$$\begin{aligned}\boldsymbol{DE} &= (a|\alpha\rangle\langle\alpha| + b|\beta\rangle\langle\beta|)|\psi\rangle\langle\psi| = \\ &= a|\alpha\rangle\langle\alpha|\psi\rangle\langle\psi| + b|\beta\rangle\langle\beta|\psi\rangle\langle\psi| = \\ &= (a\langle\alpha|\psi\rangle)|\alpha\rangle\langle\psi| + (b\langle\beta|\psi\rangle)|\beta\rangle\langle\psi|.\end{aligned}$$



Члены $\langle\alpha|\psi\rangle$ и $\langle\beta|\psi\rangle$ могут «коммутировать» с другими выражениями, так как они представляют собой просто числа, порядок же таких «объектов», как $|\alpha\rangle$ и $\langle\psi|$ необходимо тщательно соблюдать. Далее получаем (учитывая, что $z\bar{z} = |z^2|$, см. с. 412)

$$\begin{aligned}\text{СЛЕД}\,(\boldsymbol{DE}) &= (a\langle\alpha|\psi\rangle)\langle\psi|\alpha\rangle + (b\langle\beta|\psi\rangle)\langle\psi|\beta\rangle = \\ &= a|\langle\alpha|\psi\rangle|^2 + b|\langle\beta|\psi\rangle|^2.\end{aligned}$$

Напомню (см. §5.13, с. 443), что величины $|\langle\alpha|\psi\rangle|^2$ и $|\langle\beta|\psi\rangle|^2$ представляют собой *квантовые* вероятности соответствующих конечных состояний $|\alpha\rangle$ и $|\beta\rangle$, тогда как $a$ и $b$ суть *классические* вклады в полную вероятность. Таким образом, в окончательном выражении квантовые и классические вероятности оказываются смешаны.

В случае более общего измерения типа «да/нет» рассуждение в целом не изменяется, только вместо определенного выше проектора «$\boldsymbol{E}$» используется проектор более общего вида

$$\boldsymbol{E} = |\psi\rangle\langle\psi| + |\phi\rangle\langle\phi| + \ldots + |\chi\rangle\langle\chi|,$$

где $|\psi\rangle$, $|\phi\rangle$, $\ldots$, $|\chi\rangle$ — взаимно ортогональные нормированные состояния, заполняющие пространство **ДА**-состояний в гильбертовом пространстве. Как мы видим, проекторы обладают общим свойством

$$\boldsymbol{E}^2 = \boldsymbol{E}.$$

Вероятность получения ответа **ДА** при измерении, определяемом проектором $\boldsymbol{E}$, системы с матрицей плотности $\boldsymbol{D}$ равна следу $(\boldsymbol{DE})$ — в точности, как и в предыдущем примере.

Отметим важный факт: искомую вероятность можно вычислить, если нам всего-навсего известны матрица плотности и проектор, описывающий измерение. Нам не нужно знать, каким именно образом из индивидуальных состояний была составлена матрица плотности. Полная вероятность получается сама собой в виде соответствующей комбинации классических и квантовых вероятностей, а нам не приходится беспокоиться, какая ее часть откуда взялась.

Рассмотрим повнимательнее это любопытное переплетение классических и квантовых вероятностей в матрице плотности. Допустим, например, что у нас имеется частица со спином $\frac{1}{2}$,

и мы абсолютно не уверены, в каком спиновом состоянии (нормированном) она в данный момент пребывает — $|\uparrow\rangle$ или $|\downarrow\rangle$. Предположив, что соответствующие вероятности этих состояний равны $\frac{1}{2}$ и $\frac{1}{2}$, построим матрицу плотности

$$\boldsymbol{D} = \tfrac{1}{2}|\uparrow\rangle\langle\uparrow| + \tfrac{1}{2}|\downarrow\rangle\langle\downarrow|.$$

Простое вычисление показывает, что в точности *такая же* матрица плотности $\boldsymbol{D}$ получается в случае комбинации равных вероятностей ($\frac{1}{2}$ и $\frac{1}{2}$) любых других ортогональных возможностей — скажем, состояний (нормированных) $|\rightarrow\rangle$ и $|\leftarrow\rangle$, где $|\rightarrow\rangle = (|\uparrow\rangle + |\downarrow\rangle)/\sqrt{2}$ и $|\leftarrow\rangle = (|\uparrow\rangle - |\downarrow\rangle)/\sqrt{2}$:

$$\boldsymbol{D} = \tfrac{1}{2}|\rightarrow\rangle\langle\rightarrow| + \tfrac{1}{2}|\leftarrow\rangle\langle\leftarrow|.$$

Допустим, мы решили измерять спин частицы в направлении «вверх», т. е. соответствующий проектор имеет вид

$$\boldsymbol{E} = |\uparrow\rangle\langle\uparrow|.$$

Тогда для вероятности получения ответа **ДА**, согласно первому описанию, находим

$$\begin{aligned}\text{СЛЕД}\,(\boldsymbol{DE}) &= \tfrac{1}{2}|\langle\uparrow|\uparrow\rangle|^2 + \tfrac{1}{2}|\langle\downarrow|\uparrow\rangle|^2 = \\ &= \tfrac{1}{2}\times 1^2 + \tfrac{1}{2}\times 0^2 = \\ &= \tfrac{1}{2},\end{aligned}$$

где мы полагаем $\langle\uparrow|\uparrow\rangle = 1$ и $\langle\downarrow|\uparrow\rangle = 0$ (поскольку состояния нормированы и ортогональны). Согласно второму описанию, находим

$$\begin{aligned}\text{СЛЕД}\,(\boldsymbol{DE}) &= \tfrac{1}{2}|\langle\rightarrow|\uparrow\rangle|^2 + \tfrac{1}{2}|\langle\leftarrow|\uparrow\rangle|^2 = \\ &= \tfrac{1}{2}\times(1/\sqrt{2})^2 + \tfrac{1}{2}\times(1/\sqrt{2})^2 = \\ &= \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{1}{2};\end{aligned}$$



правое $|\rightarrow\rangle$ и левое $|\leftarrow\rangle$ состояния здесь не являются ни ортогональными, ни параллельными измеряемому состоянию $|\uparrow\rangle$, т. е. на деле $|\langle\rightarrow|\uparrow\rangle| = |\langle\leftarrow|\uparrow\rangle| = 1/\sqrt{2}$.

Хотя полученные вероятности оказываются одинаковыми (как, собственно, и должно быть, поскольку одинаковы матрицы плотности), физические интерпретации этих двух описаний совершенно различны. Мы согласны с тем, что физическая «реальность» любой ситуации описывается *некоторым* вполне определенным вектором состояния, однако существует классическая неопределенность в отношении того, каким окажется этот вектор в действительности. В первом предложенном описании атом находится либо в состоянии $|\uparrow\rangle$, либо в состоянии $|\downarrow\rangle$, и мы не знаем, в каком из двух. Во втором описании — либо в состоянии $|\rightarrow\rangle$, либо в состоянии $|\leftarrow\rangle$, и мы снова не знаем, в каком именно. Когда мы в первом случае выполняем измерение с целью выяснить, не находится ли атом в состоянии $|\uparrow\rangle$, мы имеем дело с самыми обычными классическими вероятностями: вероятность того, что атом находится в состоянии $|\uparrow\rangle$, совершенно очевидно равна $\frac{1}{2}$, и больше тут говорить не о чем. Когда мы задаем тот же вопрос во втором случае, измерению подвергается уже комбинация вероятностей состояний $|\rightarrow\rangle$ и $|\leftarrow\rangle$, и каждое из них вносит в полную вероятность свой классический вклад $\frac{1}{2}$, помноженный на свой же квантовомеханический вклад $\frac{1}{2}$, что дает в итоге $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$. Как можно видеть, матрица плотности ухитряется сосчитать нам верную вероятность вне зависимости от того, какие классические и квантовомеханические доли эту вероятность, по нашему предположению, составляют.

Приведенный выше пример является в некотором роде особым, поскольку так называемые «собственные значения» матрицы плотности в этом случае оказываются вырожденными (в силу того, что обе классические вероятности здесь — $\frac{1}{2}$ и $\frac{1}{2}$ — одинаковы); именно эта «особость» и позволяет нам составить более одного описания в комбинациях вероятностей ортогональных альтернатив. Впрочем, для наших рассуждений это ограничение несущественно. (А упомянул я о нем исключительно для того, чтобы избежать упреков в невежестве со стороны возможно читающих эти строки специалистов.) Всегда можно предста-

вить, что комбинация вероятностей охватывает гораздо большее число состояний, нежели просто набор взаимно ортогональных альтернатив. Например, в вышеописанной ситуации мы вполне могли бы составить очень сложные вероятностные комбинации множества возможных различных направлений оси спина. Иначе говоря, существует огромное количество совершенно различных способов представить одну и ту же матрицу плотности в виде комбинации вероятностей альтернативных состояний, и это верно для *любых* матриц плотности, а не только для тех, собственные значения которых вырожденны.

## 6.5. Матрицы плотности для ЭПР-пар

Перейдем к ситуациям, описание которых в терминах матриц плотности представляется особенно уместным — и в то же время выявляет один почти парадоксальный аспект интерпретации такой матрицы. Речь идет об ЭПР-эффектах и квантовой сцепленности. Рассмотрим физическую ситуацию, описанную в § 5.17: частица со спином 0 (в состоянии $|\Omega\rangle$) расщепляется на две частицы (каждая со спином $\frac{1}{2}$), которые разлетаются вправо и влево, удаляясь на значительное расстояние друг от друга, в результате чего выражение для их совокупного (сцепленного) состояния принимает вид:

$$|\Omega\rangle = |\mathbf{L}\uparrow\rangle|\mathbf{R}\downarrow\rangle - |\mathbf{L}\downarrow\rangle|\mathbf{R}\uparrow\rangle.$$

Предположим, что некий наблюдатель[4] имеет намерение измерить спин правой частицы с помощью некоего измерительного устройства, левая же частица успела уже удалиться на такое огромное расстояние, что добраться до нее наблюдатель не может. Как наш наблюдатель опишет состояние спина правой частицы?

Скорее всего, он весьма благоразумно воспользуется матрицей плотности

$$\boldsymbol{D} = \tfrac{1}{2}|\mathbf{R}\uparrow\rangle\langle\mathbf{R}\uparrow| + \tfrac{1}{2}|\mathbf{R}\downarrow\rangle\langle\mathbf{R}\downarrow|,$$

поскольку ничто не мешает ему вообразить, что некий другой наблюдатель — скажем, коллега, по случаю оказавшийся неподалеку от левой частицы, — решил измерить спин этой левой частицы в направлении «вверх/вниз». Узнать, какой именно результат получил упомянутый воображаемый коллега, нашему наблюдателю неоткуда. Однако он знает, что если коллега получил результат $|\mathbf{L}\uparrow\rangle$, то его собственная (правая) частица должна находиться в состоянии $|\mathbf{R}\downarrow\rangle$, если же коллега получил при измерении состояние $|\mathbf{L}\downarrow\rangle$, то правая частица должна находиться в состоянии $|\mathbf{R}\uparrow\rangle$. Нашему наблюдателю также известно (из стандартных правил квантовой теории, касающихся вероятностей, какие можно ожидать в данной ситуации), что воображаемый коллега может получить с равной вероятностью как результат $|\mathbf{L}\uparrow\rangle$, так и результат $|\mathbf{L}\downarrow\rangle$. Из всего этого наблюдатель заключает, что состояние его собственной частицы описывается комбинацией равных вероятностей ($\frac{1}{2}$ и $\frac{1}{2}$, соответственно) двух альтернатив, $|\mathbf{R}\uparrow\rangle$ и $|\mathbf{R}\downarrow\rangle$, так что матрица плотности $\boldsymbol{D}$ с его стороны действительно должна быть такой, какую мы только что записали.

Он, впрочем, может предположить, что его коллега производил измерение левой частицы в направлении «влево/вправо». В этом случае совершенно аналогичное вышеизложенному рассуждение (на сей раз опирающееся на альтернативное описание $|\Omega\rangle = |\mathbf{L}\leftarrow\rangle|\mathbf{R}\rightarrow\rangle - |\mathbf{L}\rightarrow\rangle|\mathbf{R}\leftarrow\rangle$, см. с. 454) приведет нашего наблюдателя к заключению, что спиновое состояние его собственной (правой) частицы описывается комбинацией равных вероятностей направлений оси спина «влево» и «вправо», а соответствующая матрица плотности имеет вид

$$\boldsymbol{D} = \tfrac{1}{2}|\mathbf{R}\rightarrow\rangle\langle\mathbf{R}\rightarrow| + \tfrac{1}{2}|\mathbf{R}\leftarrow\rangle\langle\mathbf{R}\leftarrow|.$$

Как мы уже видели, эти матрицы плотности в точности одинаковы, однако их *интерпретации* — как комбинаций вероятностей альтернативных состояний — существенно различаются. Совершенно не важно, какую именно интерпретацию выберет наблюдатель. Из своей матрицы плотности он получит всю возможную информацию, требуемую для вычисления вероятностей результатов измерений спина правой (и только правой) частицы. Более того, поскольку коллега является *воображаемым*, нашего наблюдателя вообще не должно волновать, выполнялось ли хоть какое-то измерение спина левой частицы. Все та же матрица плотности $\boldsymbol{D}$ скажет ему все, что можно узнать о состоянии спина

[4] См. обращение к читателю в начале книги, с. 18.

правой частицы до того, как он действительно выполнит измерение. В самом деле, уж наверное матрица плотности $\boldsymbol{D}$ определит «действительное состояние» правой частицы с гораздо большей точностью, нежели какой бы то ни было отдельный вектор состояния.

Руководствуясь подобными общими соображениями, люди порой приходят к выводу, что в определенных ситуациях матрицы плотности дают более адекватное описание квантовой «реальности», чем векторы состояния. Однако в ситуациях, подобных рассматриваемой, *это не так*. Ничто в принципе не мешает воображаемому коллеге превратиться в коллегу реального, а двум наблюдателям — передать друг другу результаты своих наблюдений. Корреляции между измерениями, выполненными одним наблюдателем, и измерениями, выполненными другим, невозможно объяснить отдельными матрицами плотности, описывающими каждая свою частицу. Для такого объяснения необходимо все сцепленное состояние целиком, в том виде, в каком оно представлено выше выражением для действительного вектора состояния $|\Omega\rangle$.

Например, если оба наблюдателя решат измерять спины своих частиц в направлении «вверх/вниз», то они неизбежно должны получить диаметрально противоположные результаты. Индивидуальные матрицы плотности такой информации не содержат. Еще более серьезное возражение: как недвусмысленно показывает теорема Белла (§ 5.4), моделировать сцепленное состояние связанной пары частиц какими бы то ни было локальными классическими методами (вроде «носков Бертлмана») *до* измерения *невозможно*. (Простая демонстрация этого факта приводится в НРК, примечание 14 после шестой главы, с. 301 — идея этой демонстрации, вообще говоря, принадлежит Стаппу [359], см. также [360]. Описан случай, когда один из наблюдателей измеряет спин своей частицы в вертикальном, «вверх/вниз», или горизонтальном, «вправо/влево», направлении, тогда как другой выбирает для измерения одно из направлений под углом в 45° к тем двум. Если заменить частицы со спином $\frac{1}{2}$ частицами со спином $\frac{3}{2}$, то такую демонстрацию можно сделать еще более убедительной, воспользовавшись магическими додекаэдрами из § 5.3, так как при этом нам не понадобятся вероятности.)



Таким образом, в данной ситуации «матричное» описание может быть признано адекватным «реальности», только если имеется какая-либо причина, *в принципе* не позволяющая выполнить (и сравнить) измерения на обоих концах системы. В обычных условиях таких причин, как правило, не существует. В условиях необычных — например, в ситуации, предложенной Стивеном Хокингом [191], где одна из частиц ЭПР-пары оказывается заключенной внутрь черной дыры, — могут появиться и более серьезные доводы в пользу матричного описания на фундаментальном уровне (что, собственно, и доказывает Хокинг). Однако такие доводы сами по себе предполагают некий серьезный пересмотр самих основ квантовой теории. Пока такого пересмотра не произошло, существенная роль матрицы плотности остается скорее практической (FAPP), нежели фундаментальной — что, впрочем, отнюдь не уменьшает ее важности.

## 6.6. FAPP-объяснение процедуры R

Теперь давайте посмотрим, какую же, в самом деле, роль играют матрицы плотности в рамках стандартного (FAPP-) подхода к объяснению «наблюдаемой» природы процедуры **R**. Идея заключается в том, что квантовая система и измерительное устройство (вместе с занимаемым ими окружением) — все три, предполагается, эволюционируют вместе в соответствии с процедурой **U** — *ведут себя так, будто* всякий раз, когда эффекты измерения оказываются нерасторжимо сцеплены с этим самым окружением, происходит процедура **R**.

Изначально квантовая система считается изолированной от окружения, однако в момент «измерения» в измерительном устройстве инициируются макроскопические эффекты, которые вскоре приводят к возникновению сцепленностей с элементами окружения, причем количество этих сцепленностей непрерывно возрастает. На этом этапе картина во многом напоминает описанную в предыдущем параграфе ЭПР-ситуацию. Квантовая система (вместе с только что сработавшим измерительным устройством) выступает в роли правой частицы, тогда как возмущенное окружение аналогично отдаленной левой частице. Физик, намеревающийся осмотреть измерительное устройство, играет роль, схожую с ролью наблюдателя, предполагающего

исследовать правую частицу. Наблюдатель не имеет доступа к каким бы то ни было измерениям, которые могли быть выполнены на левой частице; аналогично, нашему физику недоступна подробная картина возмущений, предположительно произведенных в окружении измерительным устройством. Окружение состоит из огромного количества случайным образом движущихся частиц, и можно смело утверждать, что детальная и точная информация относительно того, какому именно возмущению подверглись частицы окружения, будет безвозвратно потеряна для физика. Аналогичным образом, наблюдателю у правой частицы из предыдущего примера недоступны какие бы то ни было сведения о спине левой частицы. Как и в случае с правой частицей, состояние измерительного устройства адекватно описывается не отдельным вектором состояния, но матрицей плотности; соответственно, измерительное устройство рассматривается не как чистое, отдельно взятое квантовое состояние, но как комбинация вероятностей состояний. Согласно стандартной интерпретации, эта комбинация вероятностей дает те же вероятностно-взвешенные альтернативы, что мы получили бы в результате процедуры **R** — по крайней мере, с практической точки зрения.

Рассмотрим пример. Допустим, некий источник испускает фотон в направлении детектора. Между источником и детектором помещено полусеребрёное зеркало, после столкновения с которым фотон переходит в суперпозицию состояний

$$w|\alpha\rangle + z|\beta\rangle;$$

при этом состояние $|\alpha\rangle$ (пропущенный фотон) активирует детектор (**ДА**), а состояние $|\beta\rangle$ (отраженный фотон) никак детектора не затрагивает (**НЕТ**). Полагая все состояния нормированными, получим, в соответствии с процедурой **R**, следующие вероятности:

$$\text{вероятность ответа } \textbf{ДА} = |w|^2,$$
$$\text{вероятность ответа } \textbf{НЕТ} = |z|^2.$$

Поскольку зеркало *полупрозрачно* (как в исходном примере, рассмотренном в § 5.7, где теперешним $|\alpha\rangle$ и $|\beta\rangle$ соответствовали состояния $|\mathbf{B}\rangle$ и $i|\mathbf{C}\rangle$), каждая из этих вероятностей равна $\frac{1}{2}$, т. е. $|w| = |z| = 1/\sqrt{2}$.



Детектор находится первоначально в состоянии $|\Psi\rangle$, которое по поглощении фотона (в состоянии $|\alpha\rangle$) эволюционирует в состояние $|\Psi_{\text{Д}}\rangle$ (**ДА**), а в отсутствие поглощения фотона (в состоянии $|\beta\rangle$) — в состояние $|\Psi_{\text{Н}}\rangle$ (**НЕТ**). Если игнорировать окружение, то состояние системы на данном этапе имеет вид

$$w|\Psi_{\text{Д}}\rangle + z|\Psi_{\text{Н}}\rangle|\beta\rangle$$

(все состояния мы полагаем нормированными). Предположим, однако, что детектор, будучи макроскопическим объектом, сразу же вступает во взаимодействие с окружением, — частью такого окружения можно считать и «сбежавший» фотон (первоначально в состоянии $|\beta\rangle$), поглощенный стеной лаборатории. Как и прежде, детектор, в зависимости от того, зарегистрировал он фотон или нет, переходит в одно из своих новых состояний ($|\Psi_{\text{Д}}\rangle$ или $|\Psi_{\text{Н}}\rangle$, соответственно), однако в процессе перехода он по-разному возмущает окружение. Состояние окружения, сопутствующее состоянию детектора $|\Psi_{\text{Д}}\rangle$, обозначим через $|\Phi_{\text{Д}}\rangle$, а состояние окружения, сопутствующее состоянию детектора $|\Psi_{\text{Н}}\rangle$, — через $|\Phi_{\text{Н}}\rangle$ (эти состояния мы также полагаем нормированными, но не обязательно ортогональными). Полное состояние сцепленной системы можно записать так:

$$w|\Phi_{\text{Д}}\rangle|\Psi_{\text{Д}}\rangle + z|\Phi_{\text{Н}}\rangle|\Psi_{\text{Н}}\rangle.$$

До сих пор физик в процессе не участвовал, однако теперь он собирается осмотреть детектор, чтобы узнать, какой результат тот зафиксировал (**ДА** или **НЕТ**). Каким образом физик может оценить квантовое состояние детектора в момент, непосредственно предшествующий осмотру? Как и наблюдатель, измерявший в предыдущем параграфе спин правой частицы, наш физик резонно воспользуется матрицей плотности. Можно предположить, что никакого измерения окружения с целью выяснить, находится *оно* в состоянии $|\Phi_{\text{Д}}\rangle$ или $|\Phi_{\text{Н}}\rangle$, в действительности не проводилось — точно так же, как никто не измерял спин левой частицы в описанной выше ЭПР-паре. Соответственно, матрица плотности и в самом деле даст адекватное квантовое описание детектора.

Какова эта матрица плотности? Рассуждая стандартным образом[7] (который основывается на некоем частном способе моделирования упомянутого окружения — исходя при этом из неких не вполне обоснованных допущений, таких, например, как

допущение о несущественности корреляций ЭПР-типа), приходим к заключению, что матрица плотности в данном случае должна очень быстро принять вид, очень хорошее приближение к которому дает следующее выражение:

$$\boldsymbol{D} = a|\Psi_{\text{Д}}\rangle\langle\Psi_{\text{Д}}| + b|\Psi_{\text{Н}}\rangle\langle\Psi_{\text{Н}}|,$$

где

$$a = |w|^2 \quad \text{и} \quad b = |z|^2.$$

Эту матрицу плотности можно интерпретировать, как представление комбинации вероятностей двух альтернатив: регистрация детектором фотона (результат **ДА**) с вероятностью $|w|^2$ и отсутствие регистрации детектором фотона (результат **НЕТ**) с вероятностью $|z|^2$. Если бы имела место процедура **R**, то именно к такому результату и должен был бы прийти физик по завершении своего эксперимента — или нет?

Думаю, здесь следует проявить некоторую осторожность. Матрица плотности $\boldsymbol{D}$ и в самом деле позволяет физику вычислить необходимые ему значения вероятностей, *если предположить*, что альтернатив всего две: *либо* $|\Psi_{\text{Д}}\rangle$, *либо* $|\Psi_{\text{Н}}\rangle$. Но из наших рассуждений такое предположение никоим образом не следует. Вспомним из предыдущего параграфа, что матрицы плотности, как комбинации вероятностей состояний, допускают множество *альтернативных* интерпретаций. В частности, поскольку зеркало *полупрозрачно*, мы имеем здесь в точности такую же матрицу плотности, как и та, какую мы получили выше для частицы со спином $\frac{1}{2}$:

$$\boldsymbol{D} = \tfrac{1}{2}|\Psi_{\text{Д}}\rangle\langle\Psi_{\text{Д}}| + \tfrac{1}{2}|\Psi_{\text{Н}}\rangle\langle\Psi_{\text{Н}}|.$$

Можно записать ее иначе; скажем, так:

$$\boldsymbol{D} = \tfrac{1}{2}|\Psi_{\text{P}}\rangle\langle\Psi_{\text{P}}| + \tfrac{1}{2}|\Psi_{\text{Q}}\rangle\langle\Psi_{\text{Q}}|,$$

где $|\Psi_{\text{P}}\rangle$ и $|\Psi_{\text{Q}}\rangle$ — два других возможных ортогональных состояния детектора (что представляет собой, надо сказать, совершенную нелепость с точки зрения классической физики), причем

$$|\Psi_{\text{P}}\rangle = (|\Psi_{\text{Д}}\rangle + |\Psi_{\text{Н}}\rangle)/\sqrt{2} \quad \text{и} \quad |\Psi_{\text{Q}}\rangle = (|\Psi_{\text{Д}}\rangle - |\Psi_{\text{Н}}\rangle)/\sqrt{2}.$$



Тот факт, что наш физик полагает, будто состояние его детектора описывается матрицей плотности $\boldsymbol{D}$, никак не объясняет, *почему* он всегда обнаруживает детектор либо в состоянии **ДА** (что соответствует $|\Psi_{\text{Д}}\rangle$), либо в состоянии **НЕТ** ($|\Psi_{\text{Н}}\rangle$). Потому что совершенно такую матрицу плотности он получил бы, если состояние системы представляло собой равновесную вероятностную комбинацию, по классическим меркам, нелепостей $|\Psi_{\text{P}}\rangle$ и $|\Psi_{\text{Q}}\rangle$ (описывающих, соответственно, квантовые линейные суперпозиции «**ДА** *плюс* **НЕТ**» и «**ДА** *минус* **НЕТ**»)!

Для того, чтобы подчеркнуть физическую абсурдность состояний, подобных $|\Psi_{\text{P}}\rangle$ и $|\Psi_{\text{Q}}\rangle$, в случае макроскопического детектора, рассмотрим «измерительное устройство», состоящее из ящика и помещенной внутрь него кошки, причем ящик снабжен неким устройством, убивающим кошку, если детектор регистрирует фотон (в состоянии $|\alpha\rangle$), если же детектор ничего не регистрирует (фотон в состоянии $|\beta\rangle$), то кошка остается жива — это измерительное устройство широко известно под названием *шрёдингерова кошка* (см. § 5.1 и рис. 6.3). Результат **ДА** представляется здесь как «кошка мертва», а результат **НЕТ** — как «кошка жива». Однако из одного лишь того, что нам известно, что матрица плотности имеет вид равновесной комбинации этих двух состояний, вовсе *не* следует, что кошка либо мертва, либо жива (с равной вероятностью), так как эта же кошка может также быть (с равной вероятностью) либо «мертва плюс жива», либо «мертва минус жива»! Сама по себе матрица плотности *ничего* не говорит о том, что эти последние классически абсурдные возможности в известном нам реальном мире никогда не реализуются. Как и во «множественно мировом» подходе к объяснению **R**, нам, похоже, вновь предлагается поразмыслить над тем, какого рода состояния мы намерены позволить воспринимать обладающему сознанием наблюдателю (в данном случае, нашему «физику»). С чего мы, собственно говоря, взяли, что состояния вроде «кошка мертва плюс кошка жива» совершенно и абсолютно недоступны восприятию некоего сознательного внешнего[5] наблюдателя?

[5]Нельзя, разумеется, забывать и о сознании кошки! На эту сторону дела обратил наше самое пристальное внимание Юджин П. Вигнер, предложив свой вариант парадокса шрёдингеровой кошки [385]. «Друг Вигнера» разделяет с шрёдингеровой кошкой некоторые из ее лишений, однако в каждом из состояний суперпозиции остается в полном сознании!

Мне могут возразить, что «измерение» детектора, которое наш физик намерен произвести, состоит всего лишь в том, чтобы узнать, какой результат из двух (**ДА** или **НЕТ**) этот самый детектор зафиксировал — или, как в примере с кошкой, выяснить, мертва она или жива. (Вспомним и о наблюдателе из предыдущего параграфа, который собирался всего лишь определить, вверх направлена ось спина правой частицы или вниз.) Для такого измерения матрица плотности и в самом деле дает верные значения вероятностей, в каком бы виде мы ее ни представили. А вот тут начинаются проблемы. Почему мы должны считать *таким* измерением простой *взгляд* на кошку? В **U**-эволюции квантовой системы нет ни единого правила, запрещающего нашему сознанию в процессе «разглядывания» и, как следствие, *восприятия* квантовой системы осознавать комбинации вроде «кошка мертва плюс кошка жива». Так! Здесь мы, кажется, уже проходили. Что такое сознание? Как *на самом деле* устроен наш мозг? Ведь первой и самой очевидной причиной поисков FAPP-объяснения процедуры **R** как раз и было желание избежать *необходимости* связываться с такого рода вопросами!

Кто-то скажет: все дело в том, что мы выбрали для нашего примера нехарактерный особый случай с двумя *равными* вероятностями $\frac{1}{2}$ и $\frac{1}{2}$ (случай «вырожденных собственных значений»). Только в таких ситуациях матрица плотности допускает более одного представления в виде взвешенной вероятностной комбинации взаимно *ортогональных* альтернатив. Это ограничение *не существенно*, поскольку для интерпретации матрицы плотности как комбинации вероятностей ортогональность альтернатив непременным требованием не является. Более того, как показали в своей недавней работе Хьюстон, Йожа и Вуттерс [210], в ситуациях, подобных вышеописанным (т. е. там, где матрица плотности вводится потому, что рассматриваемая система сцеплена с какой-то другой изолированной системой), для *любой* комбинации вероятностей альтернативных состояний, выбранной вами для составления матрицы плотности, всегда найдется измерение, выполнимое в той самой изолированной системе, которое даст в точности такое же представление матрицы плотности. Как бы то ни было, одно то, что неоднозначность возникает уже в случае *равных* вероятностей, ясно показывает, что для описания *действительных* альтернативных состояний нашего детектора матричного представления недостаточно.



Итак, одно лишь знание матрицы плотности ***D*** *не дает* никаких оснований полагать, что система представляет собой вероятностную комбинацию тех самых состояний, которые эту конкретную матрицу ***D*** составляют. Точно такую же матрицу ***D*** можно получить и из множества других самых различных комбинаций состояний, бóльшая часть которых окажутся совершенно «абсурдными» с точки зрения здравого смысла. Более того, такая неоднозначность свойственна любой матрице плотности, какую ни возьми.

Стандартные рассуждения не часто заходят дальше требования «диагональности» матрицы плотности. «Диагональной», по сути, является такая матрица плотности, которую можно выразить в виде взвешенной вероятностной комбинации взаимно *ортогональных* альтернатив — точнее, не всяких альтернатив, а тех классических альтернатив, которые нас в данном случае интересуют. (Если убрать это последнее условие, то диагональными окажутся *все* матрицы плотности!) Однако мы уже убедились, что один лишь факт «выразимости» матрицы плотности в таком виде сам по себе отнюдь не является гарантией того, что детекторы *не предстанут* перед нами в какой-нибудь «абсурдной» квантовой суперпозиции состояний **ДА** и **НЕТ**.

Таким образом, вопреки всем и всяческим уверениям, стандартное рассуждение *не объясняет*, как то или иное приближенное описание **U**-эволюции в условиях неустранимого воздействия окружения порождает «иллюзию» процедуры **R**. Оно демонстрирует всего лишь, что в такой ситуации процедура **R** и **U**-эволюция могут мирно сосуществовать. Нам все еще нужно в квантовой теории место для процедуры **R**, отличное от того, что занимает **U**-эволюция (по крайней мере, пока не появится теория, жестко предписывающая, какого рода состояния способны воспринимать существа, обладающие сознанием).

Отыскание такого места само по себе важно для общей непротиворечивости квантовой теории. Однако не менее важно понять, что это сосуществование и эта непротиворечивость имеют статус скорее практического приближения (FAPP), нежели строго научный. В конце предыдущего параграфа мы говорили о том, что описание правой частицы посредством матрицы плотности является адекватным лишь в отсутствие возможности сравнения измерений, выполненных на *обоих* частицах. Если же такая возможность есть, то необходимо рассматривать полное

состояние системы с ее *квантовыми*, а не просто взвешенно-вероятностными суперпозициями. Аналогичным образом, матричное описание детектора в настоящем параграфе адекватно лишь в том случае, если отсутствует возможность детально измерить состояние окружения и сравнить результаты измерения с результатами наблюдения детектора экспериментатором. Редукция **R** может сосуществовать с эволюцией **U** исключительно при условии, что мельчайшие элементы окружения останутся недоступными измерению, а тонкие эффекты квантовой интерференции, надежно укрытые (согласно стандартной квантовой теории) невообразимой сложностью точного описания окружения, избегнут наблюдения.

Очевидно, что какая-то (и даже немалая) доля правды в стандартном объяснении есть, однако полным оно быть никак не может. Разве можем мы быть уверены в том, что в ближайшем будущем не появится какая-нибудь новая технология, с помощью которой все эти интерференционные феномены будут детально описаны? Необходимо ввести некое строгое физическое правило, определяющее, какие из экспериментов, невозможных сегодня практически, являются невозможными *в принципе*. Согласно такому правилу, должен существовать некий уровень физических процессов, получение каких бы то ни было данных об эффектах интерференции на котором невозможно в принципе. Придется, по всей видимости, постулировать некий *новый* физический феномен, благодаря которому комплексно-взвешенные суперпозиции физики квантового уровня *действительно* станут классическими альтернативами, а не просто будут считаться таковыми в FAPP-приближении. В существующем же виде FAPP-подход не дает картины действительной физической реальности. Он не может быть ничем иным, как временной полумерой в отсутствие настоящей физической теории — хотя и весьма полезной, надо сказать, полумерой, — и важно иметь это в виду, когда мы будем рассматривать выдвигаемые мною в § 6.12 предположения.

## 6.7. FAPP-объяснение правила квадратов модулей

В предыдущих трех параграфах неявно присутствовало одно далеко идущее допущение, к которому я намеренно не привлекал



излишнего внимания. *Одна лишь* необходимость такого допущения эффективно аннулирует любое предположение о том, что из **U**-эволюции можно *вывести* правило квадратов модулей для процедуры **R** — даже в FAPP-приближении. Уже самим фактом использования матрицы плотности мы неявно *допускаем*, что взвешенная вероятностная комбинация может быть описана таким объектом вполне адекватно. Уже сама уместность использования выражений вроде $|\alpha\rangle\langle\alpha|$ (которые, в свою очередь, принадлежат к виду «объект, умноженный на собственное комплексное сопряженное») определенно намекает на присутствие где-то рядом правила квадратов модулей. Правило получения значений вероятности из матрицы плотности корректно сочетает классические и квантовые вероятности только потому, что правило квадратов модулей *встроено* в саму концепцию матрицы плотности.

Хотя процесс унитарной эволюции (**U**) действительно очень хорошо стыкуется (математически) с концепциями матрицы плотности и скалярного произведения $\langle\alpha|\beta\rangle$ в гильбертовом пространстве, это вовсе *не означает*, что вычисляемые с помощью квадратов модулей величины непременно являются *вероятностями*. То есть речь снова идет о сосуществовании **R** и **U**, а не об объяснении происхождения **R** из **U**. Унитарной эволюции абсолютно ничего не известно о понятии вероятности. То, что квантовые вероятности можно вычислять с помощью этой процедуры, совершенно очевидно является *дополнительным* допущением, вне зависимости от того, каким образом мы пытаемся обосновать взаимоотношения процедур **R** и **U** — привлекая к делу множественность миров или используя FAPP-подход.

Поскольку почти все экспериментальные подтверждения, какими может похвастаться квантовая механика, основаны на предписываемой теорией процедуре вычисления вероятностей, игнорировать **R**-часть квантовой механики мы можем лишь на свой страх и риск. Редукция **R** отлична от эволюции **U** и не следует из **U**, как бы громко и часто теоретики ни уверяли нас в обратном. А раз так, то придется нам смириться с **R** как с отдельным, самостоятельным физическим процессом. Я отнюдь не настаиваю на немедленном присвоении редукции статуса отдельного, самостоятельного физического *закона*. Ничуть не сомневаюсь, что она представляет собой приближение чего-то такого, о чем мы, возможно, еще не имеем никакого представления. Рассужде-

ния в конце предыдущего параграфа недвусмысленно указывают на то, что применение **R**-процедуры в процессе измерения действительно носит приближенный характер.

Согласимся пока с тем, что необходимо искать какие-то новые объяснения, и попробуем, соблюдая должную осторожность, двинуться дальше теми тропами в неизвестное, что, возможно, еще открыты перед нами.

## 6.8. О редукции вектора состояния посредством сознания

Среди тех, кто всерьез полагает, что вектор состояния $|\psi\rangle$ описывает реальный физический мир, есть такие, кто утверждает — в противовес уповающим на эволюцию **U** на всех уровнях, т. е. приверженцам концепции множественности миров, — что нечто подобное процедуре **R** *действительно* происходит, причем происходит тогда, когда в процесс вовлекается сознание наблюдателя. Выдающийся физик Юджин Вигнер как-то даже набросал вкратце теорию такого процесса [385]. Общая идея заключается в том, что бессознательная материя — или, возможно, всего лишь неживая материя — эволюционирует в соответствии с **U**, однако как только состояние системы оказывается сцеплено с состоянием какого-либо сознательного (или просто «живого») существа, появляется нечто новое, в дело вступает некий физический процесс, приводящий к **R**, он-то и редуцирует *в действительности* состояние системы.

Не думаю, что есть необходимость формулировать предположение (следуя такой точке зрения), что сознательное существо каким-то образом приобретает способность оказывать «воздействие» на тот выбор, какой делает в этот момент Природа. Такое предположение увлекло бы нас в чрезвычайно коварные воды — насколько я могу судить, наблюдаемые факты резко противоречат любым подобного рода упрощенным заявлениям, сводящимся к тому, что сознательный волевой акт способен воздействовать на результат квантовомеханического эксперимента. Таким образом, мы не станем в рамках нашего исследования настаивать на том, что процедура **R** должна непременно требовать активного участия «свободной сознательной воли» (альтернативным точкам зрения, впрочем, уделено некоторое внимание в § 7.1).

Не сомневаюсь, что кое-кто из читателей ожидал, что идеи



подобного рода должны были привлечь на свою сторону и меня (раз уж я занимаюсь поиском связей между проблемой квантового измерения и проблемой сознания). Уверяю вас, *это не так*. В конце концов, вполне возможно, что в нашей Вселенной сознание — феномен достаточно редкий. На поверхности Земли обладающие сознанием существа встречаются в самых различных местах, однако, насколько позволяют судить имеющиеся у нас на данный момент экспериментальные свидетельства[8], в глубинах Вселенной, на расстоянии многих световых лет от нас, высокоразвитого — или какого-либо иного — сознания нет. Получается весьма странная картина: «реальная» физическая вселенная, физические объекты в которой эволюционируют так или иначе в зависимости от того, может ли их видеть, слышать или как-то иначе ощущать какой-либо из разумных обитателей этой самой вселенной.

Возьмем для примера погоду. Синоптические ситуации, развивающиеся в атмосфере любой планеты, обусловлены хаотическими физическими процессами (см. § 1.7) и, как следствие, очень чувствительны к многочисленным единичным квантовым событиям. Если в отсутствие сознания процесс **R** и вправду не происходит, тогда расплывчатое марево альтернатив квантовых суперпозиций никогда не сгустится в какую-то определенную синоптическую ситуацию. Можем ли мы и в самом деле полагать, что погода на какой-нибудь далекой планете так и пребывает в виде некоей совокупности комплексных суперпозиций бесконечного количества различных возможных вариантов (этакой полной неразберихи, не имеющей ничего общего с настоящей погодой), пока ее не воспримет своими органами чувств какое-нибудь забредшее туда случайно разумное создание, — в каковой момент, *и ни мгновением раньше*, вся эта куча суперпозиций превратится, наконец, в погоду?

Можно возразить, что с операционной точки зрения — т. е. с операционной точки зрения обладающего сознанием существа — такая «погода суперпозиций» ничем не отличается от *настоящей* неизвестной заранее погоды (FAPP!). Однако такое решение проблемы физической реальности не является, само по себе, удовлетворительным. Как мы уже видели, FAPP-подход не объясняет «реальность» на таком фундаментальном уровне, но служит лишь в качестве временной полумеры, которая позволяет в рамках современной квантовой механики объединить **U**- и **R**-

процедуры — до тех пор, по крайней мере, пока технический прогресс не заведет нас туда, где нам потребуется более точная и последовательная теория.

Словом, я предлагаю направить наши поиски решения проблем квантовой механики в какую-нибудь другую сторону. Хотя и нельзя исключить, что проблема разума окажется в конечном счете связана с проблемой квантового измерения — или U/R-парадоксом квантовой механики, — сознание само по себе (в том виде, в каком мы представляем его себе сейчас) не способно, по моему глубокому убеждению, разрешить внутренние физические конфликты квантовой теории. Думаю, что мы должны обратиться к проблеме квантового измерения и решить ее *прежде*, чем можно будет ожидать какого-либо реального прогресса в объяснении сознания в терминах физических процессов — причем решать эту проблему следует исключительно *физическими* средствами. Когда у нас появится удовлетворительное решение, мы, возможно, окажемся в лучшем положении для поиска ответов на загадку сознания. Я считаю, что решение проблемы квантового измерения является *необходимым условием* для понимания работы разума, но *никогда* не утверждал, что это одна и та же проблема. Проблема разума неизмеримо сложнее проблемы измерения!

## 6.9. А теперь попробуем принять $|\psi\rangle$ действительно всерьез

Как выяснилось, те точки зрения, что на данный момент претендуют на серьезное отношение к квантовому описанию мира, *в действительности* всерьез его не принимают. Возможно, квантовый формализм слишком нам чужд, чтобы его можно было с легкостью принимать всерьез, и большинство физиков опасается чересчур сильно в него углубляться. Ведь кроме вектора состояния $|\psi\rangle$, эволюционирующего согласно **U**, пока система остается на квантовом уровне, нам приходится здесь иметь дело с крайне неприятным, дискретным и вероятностным, действием процедуры **R**, которое, по всей видимости, вызывает дискретные «скачки» вектора $|\psi\rangle$, когда квантовые эффекты переходят на классический уровень. Таким образом, если мы намерены предположить, что вектор $|\psi\rangle$ описывает *реальность*, то необходимо признать физически реальными и эти *скачки*, как бы неуютно мы себя в этой связи ни чувствовали. Впрочем, если мы и впрямь



принимаем реальность описания в терминах квантового вектора состояния *настолько* всерьез, то нам следует быть готовыми к внесению в существующие правила квантовой теории некоторых (предпочтительно очень тонких) изменений, поскольку действие эволюции **U**, строго говоря, несовместимо с процедурой **R** и для того, чтобы прикрыть зияющие провалы между описаниями квантового и классического уровней поведения, нам предстоит проделать некоторую деликатную «бумажную работу».

Надо сказать, что за последние годы уже было предпринято несколько попыток построить на основании этих соображений нетрадиционную непротиворечивую теорию. В 1966 году ученые венгерской школы под руководством Карольхази (Будапешт) представили [216] точку зрения, согласно которой реальный физический феномен **R**-процедуры обусловлен гравитационными эффектами (см. также [227]). Следуя несколько иной линии рассуждения, Филип Перл из Гамильтон-колледжа (Клинтон, шт. Нью-Йорк, США) выдвинул в 1976 году [284] негравитационную теорию, в которой **R** также фигурировала в качестве реального физического феномена. Позднее, в 1986 году, Джанкарло Гирарди, Альберто Римини и Туллио Вебер предложили новый интересный подход к решению проблемы; подход этот получил весьма положительную оценку самого Джона Белла, вследствие чего не заставили себя ждать многочисленные дальнейшие доработки и усовершенствования оригинальной идеи другими исследователями[(9)].

Прежде чем мы перейдем в следующих параграфах к изложению моей собственной точки зрения на предмет, немало позаимствовавшей из схемы Гирарди—Римини—Вебера (ГРВ-схемы), будет полезно ознакомиться вкратце с собственно оригиналом. Основная идея состоит в том, что вектор состояния $|\psi\rangle$ предполагается реальным, а **U**-процедуры — в основном точными. Тогда, согласно уравнению Шрёдингера, волновая функция отдельной, изначально локализованной свободной частицы стремится с течением времени распространиться во всех направлениях в пространстве (см. рис. 6.1). (Вспомним, что волновая функция частицы определяет комплексные весовые коэффициенты для различных возможных местоположений этой самой частицы. Графики на рис. 6.1 мы можем рассматривать как схематические описания поведения вещественных частей этих весовых коэффициентов.) Таким образом, со временем частица становит-

ся все менее и менее локализованной. Новым в ГРВ-схеме является допущение, что существует некоторая очень малая вероятность того, что волновая функция частицы внезапно умножится на функцию с выраженным максимумом (так называемую *гауссову* функцию) и известным размахом, определяемым некоторым параметром $\sigma$. Это событие схематически показано на рис. 6.2. При этом происходит мгновенная локализация волновой функции частицы, после чего функция вновь начинает «расползаться» вширь. Вероятность того, что пик гауссовой функции придется на то или иное конкретное местоположение частицы, пропорциональна квадрату модуля значения ее волновой функции в этой точке. Таким образом достигается совместимость со стандартным «правилом квадратов модулей» квантовой теории.

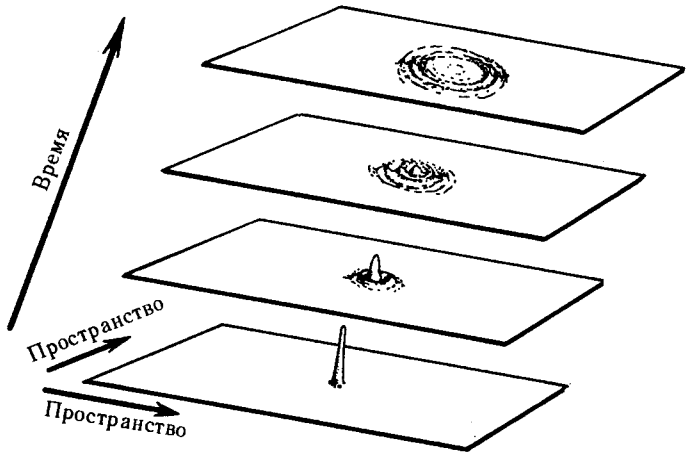



Рис. 6.1. Шрёдингерова эволюция волновой функции частицы во времени: первоначально функция плотно локализована в одной точке, а затем распространяется во всех направлениях в пространстве.

Как часто происходит подобная процедура? Предполагается, что приблизительно раз в сто миллионов ($10^8$) лет. Обозначим этот период времени буквой $T$. Тогда вероятность того, что такая редукция состояния случится с частицей в течение, скажем, одной секунды, составит менее $10^{-15}$ (поскольку секунд в году около $3 \times 10^7$). Таким образом, в случае единичной частицы никто бы



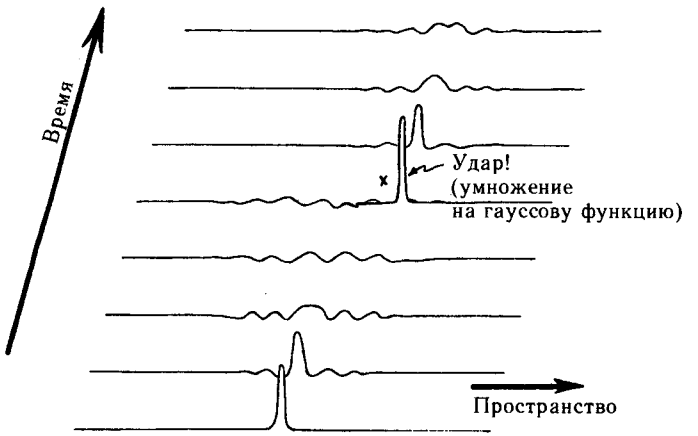



Рис. 6.2. В первоначальной схеме Гирарди – Римини – Вебера (ГРВ-схеме) волновая функция большую часть времени эволюционирует согласно стандартной шрёдингеровой **U**-эволюции, однако приблизительно раз в $10^8$ лет (на одну частицу) состояние частицы претерпевает своего рода «удар», при котором волновая функция частицы умножается на гауссову функцию с выраженным максимумом — ГРВ-интерпретация процедуры **R**.

ничего и не заметил. А теперь представьте себе, что у нас имеется некий достаточно большой объект, каждая из частиц которого подвергается той же самой процедуре. Если наш объект содержит порядка $10^{25}$ частиц (примерно столько умещается в небольших размеров мыши), то вероятность того, что *какая-либо* из его частиц испытает такого рода «удар», чрезвычайно возрастает, и можно ожидать, что удары внутри объекта будут происходить с интервалом приблизительно в $10^{-10}$ секунд. Каждый такой удар будет воздействовать на состояние объекта в целом, поскольку предполагается, что состояние каждой конкретной частицы, испытавшей удар, сцеплено с состояниями остальных частиц объекта.

Попробуем применить такой подход к *шрёдингеровой кошке*[10]. Этот парадокс — главная, в сущности, **X**-загадка квантовой теории — возникает, когда макроскопический объект (например, кошка) помещается в квантовую линейную суперпозицию

двух очевидно различных состояний, скажем, «кошка жива» и «кошка мертва» (см. также §§ 5.1 и 6.6). В квантовомеханическом смысле в такой суперпозиции ничего необычного нет, однако если рассматривать результирующую ситуацию как феномен окружающего нас с вами *реального* мира, то она представляется крайне невероятной, — что Шрёдингер неустанно подчеркивал (отдельные «$|\psi\rangle$-реалисты», впрочем, Шрёдингеру не поверили и решили отыскать-таки разгадку, обратившись кто к множественности миров, кто к редукции состоянии посредством сознания, кто еще куда; см., например, §§ 6.2 и 6.8). Для построения модели шрёдингеровой кошки нам необходимо лишь некое подходящее квантовое событие, вызывающее макроскопический эффект, — по сути, *измерение*. Например, единичный фотон, испущенный источником и либо отраженный от полупрозрачного зеркала, либо прошедший сквозь него (см. § 5.7). Допустим, что пропущенная часть волновой функции фотона вызывает срабатывание детектора, который соединен с неким устройством, убивающим кошку, тогда как отраженная часть минует детектор, и кошка остается жива (см. рис. 6.3). Как и в приведенном выше рассуждении (§ 6.6) результатом будет сцепленное состояние, одна часть которого включает в себя мертвую кошку, а другая — живую кошку и вылетающий из системы фотон. Обе возможности входят в вектор состояния *одновременно* до тех пор, пока не произойдет редукция (**R**). Вот эта вот загадка «измерения» и составляет центральную **X**-загадку квантовой теории.

В схеме же ГРВ одна из частиц объекта «кошачьих» размеров (что-то около $10^{27}$ ядерных частиц) почти мгновенно «ударяется» гауссовой функцией (см. рис. 6.2), и, поскольку состояние любой отдельной частицы сцеплено с состояниями всех остальных частиц кошки, редукция состояния этой частицы «увлекает» за собой всю кошку, каковая тут же оказывается либо живой, либо мертвой. Таким образом разрешается **X**-загадка шрёдингеровой кошки — и проблемы измерения вообще.

Схема чрезвычайно остроумна, однако страдает некоторой нарочитостью. Нигде больше в физике вы не найдете никаких указаний на подобные процессы, сами же предполагаемые значения $T$ и $\sigma$ были просто «взяты с потолка», с тем чтобы получить «приемлемые» результаты. (В 1989 году Диози предложил [92] схему, напоминающую схему ГРВ, только параметры $T$ и $\sigma$ здесь уже связываются с ньютоновской гравитационной постоянной $G$.



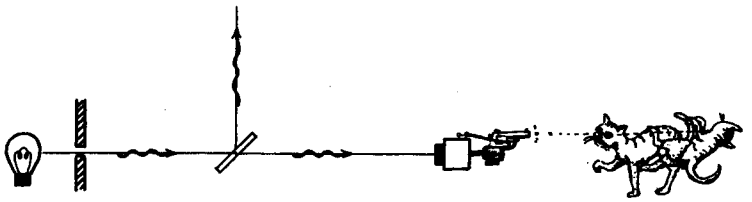


Рис. 6.3. *Шрёдингерова кошка*. Соответствующее квантовое состояние представляет собой линейную суперпозицию отраженного и пропущенного фотона. Пропущенный компонент вызывает срабатывание устройства, которое убивает кошку; иначе говоря, согласно **U**-эволюции, кошка существует в суперпозиции жизни и смерти. В ГРВ-схеме ситуация разрешается, поскольку составляющие кошку частицы почти мгновенно начинают испытывать «удары», первый же из которых локализует состояние кошки — и кошка оказывается *либо* жива, *либо* мертва.

С идеями Диози перекликаются те, что будут изложены в следующем параграфе.) Более серьезным возражением против подобного рода схем является то, что они подразумевают нарушение принципа *сохранения энергии* (пусть и незначительное). Подробнее эту важную проблему мы обсудим в § 6.12.

## 6.10. Гравитационная редукция вектора состояния

Есть веские причины[6] подозревать, что модификация квантовой теории — необходимая, если мы намерены выдать ту или иную форму **R** за *реальный* физический процесс, — должна самым серьезным образом задействовать эффекты *гравитации*. Некоторые из этих причин связаны с тем фактом, что сама структура стандартной квантовой теории очень плохо уживается с концепцией искривленного пространства, которая является неотъемлемой частью эйнштейновской теории гравитации. Даже такие

[6] Эти причины я уже изложил весьма подробно в НРК (главы 7 и 8) и не вижу необходимости повторять свои рассуждения здесь. Достаточно будет сказать, что все они до сих пор остаются в силе — хотя критерий редукции из § 6.12 существенно отличается от того, что был представлен в НРК (на с. 367–371).

понятия, как энергия и время — понятия, участвующие в фундаментальных процедурах квантовой теории, — невозможно точно определить во вполне общем гравитационном контексте, сохранив совместимость с самыми обычными требованиями стандартной квантовой теории. Вспомним также об эффекте «наклона» световых конусов (§ 4.4), уникальном свойстве физического феномена гравитации. Можно, таким образом, предположить, что ожидаемая модификация основных принципов квантовой теории явится результатом ее закономерного (и окончательного) объединения с общей теорией относительности Эйнштейна.

Впрочем, большинство физиков, похоже, не склонны допускать возможность того, что для обеспечения успеха подобного союза модификации следует подвергнуть именно *квантовую* теорию. Модификации, по их мнению, требует сама теория Эйнштейна. Они указывают (и, надо сказать, не без оснований) на то, что в классической общей теории относительности хватает и своих проблем, поскольку она предполагает существование пространственно-временны́х сингулярностей — таких, например, как черные дыры и собственно Большой Взрыв, — где кривизна пространства достигает бесконечности, а сами понятия пространства и времени вообще теряют смысл (см. НРК, гл. 7). Я нисколько не сомневаюсь, что в процессе слияния двух теорий нам предстоит модифицировать и общую теорию относительности. Равно как не вызывает сомнения и то, что такая модификация поможет нам лучше понять, что же *в действительности* происходит в тех областях, которые мы сегодня называем «сингулярностями». Но это отнюдь не освобождает квантовую теорию от необходимости пересмотра. В § 4.5 мы могли убедиться, что общая теория относительности исключительно точна — ничуть не менее точна, чем та же квантовая теория. Когда мы, наконец, сумеем должным образом эти две теории объединить, бóльшая часть физических основ как теории Эйнштейна, так и квантовой теории непременно войдет в полученную в результате общую теорию, причем в неизменном виде.

Тем не менее, многие из тех, кто мог бы, в принципе, с вышесказанным согласиться, все не унимаются: соответствующие масштабы длины, в которых способна действовать *какая бы то ни было* форма квантовой гравитации, совершенно не годятся для решения проблемы квантового измерения. В самом деле, масштаб длины, характерный для квантовой гравитации (так называемая *планковская длина*), составляет $10^{-33}$ см, что даже меньше (где-то на 20 порядков) диаметра ядерной частицы. Нас строго спрашивают, каким же это таким образом физические взаимодействия на столь крохотных расстояниях могут пролить свет на проблему измерения, которая как-никак имеет дело с феноменами уровня, пограничного (по меньшей мере) с макроскопическим. Все эти вопросы и возражения вызваны только и исключительно неверным пониманием применения идеи квантовой гравитации к данному случаю. Масштаб $10^{-33}$ см имеет к проблеме квантового измерения самое непосредственное отношение, но не в том смысле, какой первым делом приходит в голову.



Рассмотрим ситуацию, аналогичную той, в какой оказалась шрёдингерова кошка, — аналогичную тем, что здесь мы также попытаемся получить состояние линейной суперпозиции двух макроскопически различимых альтернатив. Пример такой ситуации представлен на рис. 6.4: фотон падает на полупрозрачное зеркало и оказывается в результате в состоянии линейной суперпозиции пропущенного и отраженного состояний. Пропущенная часть волновой функции фотона активирует (или способна активировать) устройство, которое перемещает некий макроскопический массивный сферический объект (не кошку) из одного пространственного положения в другое. До тех пор, пока действует шрёдингерова эволюция **U**, «местоположение» объекта определяется квантовой суперпозицией состояний «объект на прежнем месте» и «объект переместился на новое место». Как только в действие вступает редукция **R**, рассматриваемая как реальный физический процесс, объект скачкообразно занимает либо одно положение, либо другое — т. е. происходит собственно «измерение». Идея заключается в том, что, как и в ГРВ-теории, процесс этот является целиком и полностью объективным и физическим и происходит всякий раз, когда масса объекта (или расстояние, на которое он перемещается) достигает достаточной величины. (В частности, этот процесс никоим образом не зависит от того, «воспринимает» ли перемещение объекта или отсутствие такового некое случайно оказавшееся поблизости обладающее сознанием существо.) Допустим, что *устройство*, которое регистрирует прибытие фотона и перемещает объект, само по себе достаточно мало́ и может рассматриваться исключительно квантовомеханически, а измерению подвергается только лишь сферический массивный объект. В крайнем случае, мы можем просто-

напросто вообразить, что объект установлен настолько неустойчиво, что силы удара одного-единственного фотона вполне достаточно для того, чтобы вызвать значительное его смещение.

Рис. 6.4. Оставив в покое кошку, выберем в качестве предмета измерения движение сферического макроскопического объекта. Насколько велик или массивен должен быть объект, или насколько далеко он должен переместиться для того, чтобы произошла редукция **R**?

Применив стандартные **U**-процедуры квантовой механики, находим, что состояние фотона после его столкновения с зеркалом складывается из двух компонентов в очень разных положениях. Один из компонентов оказывается далее сцеплен с устройством и в конечном счете со сферическим объектом, т. е. получаем квантовое состояние, представляющее собой линейную суперпозицию двух различных местоположений объекта. Объект имеет собственное гравитационное поле, которое также следует учесть в этой суперпозиции. Таким образом, в состояние добавляется суперпозиция двух различных гравитационных полей. Согласно теории Эйнштейна, отсюда следует, что наша суперпозиция охватывает две различные пространственно-временные геометрии! Закономерно возникает вопрос: существует ли точка, в которой эти две геометрии расходятся настолько, что становятся неприменимыми правила квантовой механики, в результате чего Природа прекращает «укладывать» в суперпозицию две разные геометрии и выбирает из них какую-то одну — т. е. *физически* осуществляет некую **R**-подобную процедуру редукции?

Дело в том, что мы не имеем ни малейшего понятия, как поступать с линейными суперпозициями состояний в тех случаях, когда эти самые состояния включают в себя различные пространственно-временны́е геометрии. На этот счет «стандартная теория» может порадовать нас лишь фундаментальным пробелом: в случае существенного различия между пространственно-временны́ми геометриями мы не располагаем никакими абсолютными средствами, позволяющими сопоставить точку одной геометрии какой-либо определенной точке другой (поскольку эти геометрии представляют собой строго *разделенные* пространства), в связи с чем сама идея возможности построения суперпозиции *материальных* состояний в таких раздельных пространствах представляется крайне сомнительной.



Осталось только выяснить, когда же две геометрии *становятся* «существенно различными». Вот *тут-то* на сцену и выходит планковская длина $10^{-33}$ см. Рассуждение выглядит приблизительно так: для того чтобы произошла редукция, масштаб различия между этими геометриями должен составлять, в некотором подходящем смысле, величину порядка $10^{-33}$ см или более. Можно попробовать, например, представить себе (см. рис. 6.5), что две различные геометрии стремятся, как правило, слиться в одну, однако когда мера их различия становится для такого масштаба слишком велика, происходит редукция **R** — и вместо того, чтобы поддерживать суперпозицию, предполагаемую эволюцией **U**, Природа вынуждена выбирать какую-то одну из имеющихся геометрий.

Какой масштаб массы (или расстояния, на которое переместится объект) соответствует столь малому изменению в геометрии пространства-времени? Вообще говоря, именно благодаря малости гравитационных эффектов масштаб этот оказывается величиной довольно-таки значительной и вполне годится на роль демаркационной линии между квантовым и классическим уровнями. Для придания картине большей наглядности, необходимо еще сказать несколько слов о так называемых *абсолютных* (или *планковских*) *единицах*.

## 6.11. Абсолютные единицы

Идея (первоначально[7] предложенная Максом Планком (1906) [308] и доведенная до блеска Джоном А. Уилером (1975)

[7] Двадцатью пятью годами раньше очень похожую идею выдвинул ирландский физик Джордж Джонстон Стоуни [362]; правда, в качестве одной из основных единиц он выбрал не постоянную Планка (о существовании которой тогда никто и не подозревал), а заряд электрона. (На это мое упущение мне указал Джон Барроу, за что я ему чрезвычайно благодарен.)

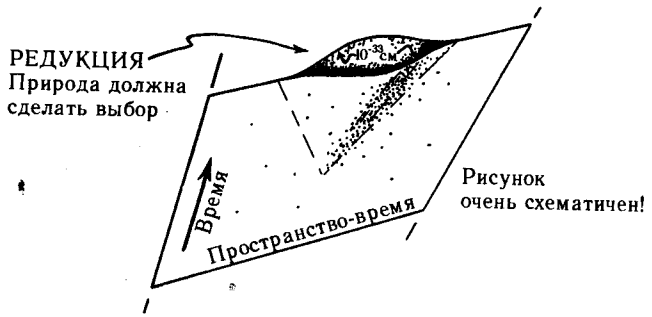



Рис. 6.5. Планковская длина $10^{-33}$ см и редукция квантового состояния. Идея заключается примерно в следующем: редукция происходит тогда, когда разница между состояниями в суперпозиции подразумевает перемещение достаточно большой массы на достаточно большое расстояние (такой массы и на такое расстояние, что различие между соответствующими пространствами–временами составляет величину порядка $10^{-33}$ см).

[383]) заключается в том, что три наиболее фундаментальные постоянные Вселенной — скорость света $c$, постоянная Планка (разделенная на $2\pi$) $\hbar$ и ньютоновская гравитационная постоянная $G$ — используются в качестве единиц для преобразования всех физических мер в чистые (безразмерные) числа. Для этого единицы длины, массы и времени необходимо выбрать таким образом, чтобы каждая из трех вышеупомянутых постоянных стала равна единице:

$$c = 1, \quad \hbar = 1, \quad G = 1.$$

Планковская длина $10^{-33}$ см, которая в обычных единицах выражается в виде $(G\hbar/c^3)^{1/2}$, принимает при этом простое значение 1 и оказывается, таким образом, абсолютной единицей *длины*. Соответствующая единица *времени*, т. е. время, за которое свет пройдет расстояние, равное планковской длине, называется *планковским временем* $((G\hbar/c^5)^{1/2})$ и равна приблизительно $10^{-43}$ секунд. Существует также абсолютная единица *массы*, так называемая *планковская масса* $((\hbar c/G)^{1/2})$, равная $2 \times 10^{-5}$ г — масса, чрезвычайно большая с точки зрения масштаба обычных квантовых феноменов, однако весьма незначительная



в нашем повседневном понимании — примерно столько весит блоха.

Понятно, что в классическом мире единицы эти не очень удобны — за исключением, разве что, планковской массы, — однако они оказываются как нельзя более полезными при рассмотрении эффектов, предположительно связанных с квантовой гравитацией. Ниже приведены некоторые из наиболее значимых физических величин, выраженные в абсолютных единицах (очень приблизительно):

$$\begin{aligned}
\text{секунда} &= 1,9 \times 10^{43} \\
\text{сутки} &= 1,6 \times 10^{48} \\
\text{год} &= 5,9 \times 10^{50} \\
\text{метр} &= 6,3 \times 10^{34} \\
\text{сантиметр} &= 6,3 \times 10^{32} \\
\text{микрон} &= 6,3 \times 10^{28} \\
\text{ферми («радиус сильного взаимодействия»)} &= 6,3 \times 10^{19} \\
\text{масса нуклона} &= 7,8 \times 10^{-20} \\
\text{грамм} &= 4,7 \times 10^{4} \\
\text{эрг} &= 5,2 \times 10^{-17} \\
\text{кельвин} &= 4 \times 10^{-33} \\
\text{плотность воды} &= 1,9 \times 10^{-94}.
\end{aligned}$$

## 6.12. Новый критерий

В этом параграфе я сформулирую новый критерий[11] гравитационной редукции вектора состояния, существенно отличный от того, что был предложен в НРК, но близкий к некоторым идеям, высказанным в последнее время Диози и другими учеными. *Причины*, побудившие меня к поискам связи между **R**-процедурой и гравитацией, остаются в силе, однако моя теперешняя гипотеза получила с тех пор дополнительную теоретическую поддержку с другой стороны. Более того, мне удалось избавиться от некоторых концептуальных проблем, присущих прежнему варианту, и сделать его более удобным для применения. В НРК я

предлагал отыскать критерий, который позволял бы определить, когда два состояния (каждое со своим гравитационным полем — т. е. пространством-временем) оказываются слишком различными для того, чтобы продолжать сосуществовать в квантовой линейной суперпозиции. Соответственно, на этом этапе должна была происходить редукция **R**. Нынешняя идея несколько отличается от прежней. Мы больше не ищем некую абсолютную меру гравитационной разницы между состояниями, чтобы выяснить с ее помощью, в какой момент состояния разойдутся настолько, что суперпозиция станет невозможна. Вместо этого, мы рассматриваем суперпозицию сколь угодно разных состояний как *нестабильную* — в том смысле, в каком нестабильно, например, ядро урана — и вводим величину *скорости* редукции вектора состояния, каковая скорость определяется как раз степенью разности состояний. Чем больше разность, тем выше скорость редукции.

Для наглядности применим новый критерий сначала к конкретной ситуации, описанной в §6.10, хотя его несложно обобщить и на многие другие случаи. Нас, в частности, интересует *энергия*, необходимая в упомянутой ситуации для того, чтобы сдвинуть одну копию объекта относительно другой, с учетом лишь *гравитационных* эффектов. Итак, мы представляем себе, что два объекта (две массы) первоначально занимают один и тот же объем пространства (см. рис. 6.6); затем одна копия объекта начинает медленно удаляться от другой, уменьшая по мере движения степень взаимопроникновения, пока, наконец, не произойдет полное их разделение, т. е., в контексте рассматриваемой ситуации, пока не будет достигнута суперпозиция состояний. Взяв величину, обратную затраченной на эту операцию гравитационной энергии (в абсолютных единицах[8]), мы получим приближенное время (также в абсолютных единицах), по истечении которого произойдет редукция состояния, в результате которой объект из состояния суперпозиции самопроизвольно и скачкообразно перейдет в то или иное локализованное состояние.

Если в качестве объекта был выбран шар с массой $m$ и

[8]Ничто, впрочем, не мешает нам выразить время редукции в более привычных, нежели введенные выше абсолютные, единицах. В этом случае время редукции определяется просто как $\hbar/E$, где $E$ — все та же гравитационная энергия разделения, а $\hbar$ — единственная постоянная, которая нам понадобится. То обстоятельство, что в выражении никак не участвует скорость света $c$, наводит на мысль о целесообразности рассмотрения теории «ньютоновской» модели такого рода (см., напр., [50]).



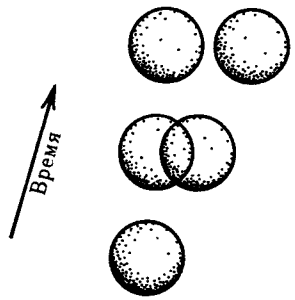



Рис. 6.6. Для того чтобы найти время редукции $\hbar/E$, представим себе объект в виде двух расходящихся копий и вычислим энергию $E$, затрачиваемую на такое расхождение, учитывая лишь гравитационное притяжение объектов.

радиусом $a$, то для энергии мы получим величину порядка $m^2/a$. Вообще говоря, действительное значение энергии зависит еще и от того, на какое расстояние перемещается объект, однако в данном случае это расстояние очень незначительно, поскольку в окончательной конфигурации две копии объекта расходятся лишь настолько, чтобы не перекрывать друг друга. Дополнительная энергия, необходимая для перемещения объекта от точки касания на любое расстояние (вплоть до бесконечности), есть величина того же порядка (коэффициент $\frac{5}{7}$), что и энергия, затрачиваемая на перемещение от полного взаимоперекрытия до точки касания. Таким образом, пока нас интересует лишь порядок величины; вкладом в общую энергию, вносимым расхождением копий объекта уже после разделения, можно пренебречь, коль скоро разделение (по большей части) таки состоялось. Согласно такой схеме, время редукции составит величину порядка

$$\frac{a}{m^2}$$

(в абсолютных единицах) или, очень приближенно,

$$\frac{1}{20\rho^2 a^5},$$

где $\rho$ — плотность объекта. То есть в случае объекта обычной

плотности (скажем, капли воды) время редукции примерно равно $10^{186}/a^5$.

В определенных простых ситуациях эта схема дает вполне «приемлемые» значения. Возьмем, например, нуклон (протон или нейтрон): если $a$ — это «радиус сильного взаимодействия» $10^{-13}$ см, что в абсолютных единицах составляет почти $10^{20}$, а масса $m$ приблизительно равна $10^{19}$, то время редукции будет что-то около $10^{58}$, т. е. более десяти миллионов лет. То, что это время велико, обнадеживает, поскольку на отдельных нейтронах эффекты квантовой интерференции наблюдались экспериментально(12). Получи мы очень малое время редукции, наши рассуждения вошли бы в противоречие с результатами этих наблюдений.

Объекты более «макроскопические», скажем, мельчайшие водяные капли радиуса $10^{-5}$ см, дадут время редукции порядка нескольких часов. Если увеличить радиус до $10^{-4}$ см (1 микрон), то время редукции уменьшится до приблизительно двенадцатой доли секунды; при радиусе $10^{-3}$ см время редукции составит менее одной миллионной секунды. В общем случае, при рассмотрении объекта в суперпозиции двух пространственно разделенных состояний мы просто определяем, какую энергию необходимо затратить на такое разделение, учитывая при этом лишь гравитационное взаимодействие между двумя «участниками» суперпозиции. Величина, обратная этой энергии, представляет собой нечто вроде «периода полураспада» суперпозиции состояний. Чем больше энергия, тем меньше время, в течение которого может существовать суперпозиция.

В реальной экспериментальной ситуации чрезвычайно сложно добиться того, чтобы объекты в квантовой суперпозиции не оказывали возмущающего воздействия на вещество окружения (образуя тем самым сцепленное с ним состояние), вследствие чего приходится учитывать и гравитационные эффекты, связанные с окружением. Такая необходимость возникает даже в тех случаях, когда возмущение не вызывает значительного макроскопического перемещения масс в окружении. Существенными могут оказаться даже самые незначительные перемещения отдельных частиц — хотя здесь для редукции обычно требуются несколько бóльшие общие массы, нежели в случае перемещения макроскопического «объекта».

Для того, чтобы наглядно продемонстрировать, какой эффект возмущение такого рода может оказать на предлагаемую схему, заменим перемещающее устройство в вышеописанной идеализированной экспериментальной ситуации неким объемом жидкости, которая просто-напросто *поглощает* фотон, если тот ухитряется пройти сквозь зеркало (см. рис. 6.7), так что теперь роль «окружения» отводится уже самому объекту. Вместо линейной суперпозиции двух состояний, различных на макроскопическом уровне в силу того, что одна копия объекта вся целиком перемещается относительно другой, мы теперь рассматриваем всего лишь различие между двумя конфигурациями взаимного расположения атомов, причем смещение одной конфигурации относительно другой носит случайный характер. Можно ожидать, что для объема обычной жидкости радиуса $a$ мы получим время редукции порядка $10^{130}/a^3$ (точная величина будет зависеть до некоторой степени от первоначальных допущений), что существенно отличается от $10^{186}/a^5$, времени редукции в опыте со взаимным перемещением объектов. То есть редукция в случае перемещения объектов целиком требует меньших масс, нежели редукция в случае возмущения атомных конфигураций. Тем не менее, в соответствии с нашей схемой редукция произойдет *и здесь*, при полном отсутствии какого бы то ни было макроскопического движения.



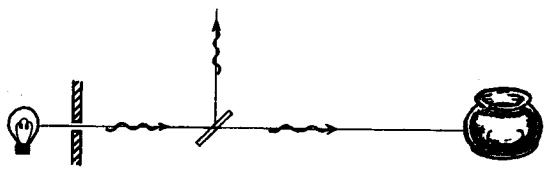


Рис. 6.7. Предположим, что пропущенный сквозь зеркало фотон не перемещает сферический объект, а всего лишь поглощается неким объемом жидкости.

В §5.8 при обсуждении квантовой интерференции мы рассматривали экспериментальную установку с материальным препятствием, перехватывающим фотонный луч. Простого *поглощения* — или даже потенциальной возможности поглощения — фотона таким препятствием вполне достаточно для редукции **R**, несмотря на то, что при этом не происходит ничего макроскопи-

ческого, что можно было бы реально наблюдать. Иначе говоря, достаточно сильное возмущение окружения, *сцепленного* с рассматриваемой системой, само по себе способно вызвать **R**, что отсылает нас к более традиционным FAPP-процедурам.

В самом деле, практически любой реальный процесс измерения почти наверняка сопровождается возмущением большого количества микроскопических частиц окружения. Согласно выдвигаемым здесь предположениям, часто доминантным эффектом оказывается именно это *возмущение*, а вовсе не макроскопическое движение массивных объектов, как в описанной выше ситуации с перемещением шара. Если эксперимент не подразумевает особо тщательного контроля за окружением, любое макроскопическое перемещение макроскопического же объекта весьма существенно возмущает окружающую среду, и вполне возможно, что именно время редукции *окружения* — величина порядка $10^{130}/b^3$, где буквой $b$ обозначен радиус области окружения, сцепленной с рассматриваемым объектом (плотность окружения принимается равной плотности воды) — оказывается в данном случае доминирующим (т. е. гораздо меньшим, нежели время редукции $10^{186}/a^5$, характерное для собственно объекта). Например, если радиус $b$ возмущенного окружения составляет всего лишь десятую долю миллиметра, то только по одной этой причине время редукции сократится до миллионной доли секунды.

Такая картина во многом близка к традиционному описанию, о котором мы говорили в §6.6, однако теперь у нас имеется вполне *определенный* критерий, позволяющий точно сказать, когда действительно происходит редукция в данном окружении. Вспомним возражения, высказанные в §6.6 против допущения, что традиционный FAPP-подход адекватно описывает действительную физическую реальность. С введением такого критерия эти возражения больше не имеют силы. Как только окружение подвергается достаточно сильному возмущению, в этом окружении очень быстро происходит (*действительно* происходит) редукция — каковая редукция незамедлительно сопровождается редукцией в любом «измерительном устройстве», с каким окружение на тот момент сцеплено. Редукция эта принципиально необратима, и восстановить первоначальное сцепленное состояние невозможно, какие бы сногсшибательные достижения технического прогресса мы себе ни вообразили. Соответственно, не возникает и противоречия с тем, что реальные измерительные устройства неизменно регистрирует *либо* **ДА**, *либо* **НЕТ** — в предлагаемой картине они делают в точности то же самое.



Мне думается, что подобного рода описание может оказаться весьма полезным при изучении различных биологических процессов; в частности, с его помощью можно вполне правдоподобно объяснить, почему биологические структуры размерами много меньше микрона часто способны на самое что ни на есть классическое поведение. Поскольку биологическая система очень тесно сцеплена со своим окружением описанном выше образом, ее *собственное* состояние непрерывно подвергается редукции вследствие столь же непрерывной редукции этого самого *окружения*. С другой стороны, можно предположить, что по какой-то причине биологическая система может «предпочесть», чтобы в тех или иных обстоятельствах ее состояние не редуцировалось в течение некоторого длительного промежутка времени. В этом случае системе необходимо найти какой-нибудь эффективный способ изоляции от окружающего ее вещества. К этим соображениям мы в дальнейшем еще вернемся (§7.5).

Следует особо подчеркнуть, что энергия, определяющая время существования суперпозиции состояний, представляет собой *разницу* энергий, а не *общую* (массу-)энергию всей системы как целого. Таким образом, в тех случаях, когда перемещаемый объект хотя и велик, но передвигается на небольшое расстояние (и если он к тому же обладает еще и кристаллической структурой, т. е. составляющие его отдельные атомы не склонны к случайным блужданиям), квантовые суперпозиции могут сохраняться в течение довольно долгого времени. Такой объект может быть гораздо больше, чем рассматриваемые выше водяные капли. Поблизости вполне «безнаказанно» могут находиться и другие, гораздо большие массы — при условии, что они не сцеплены сколько-нибудь существенно с нашей суперпозицией состояний. (Эти соображения играют важную роль при конструировании различных твердотельных устройств, таких, например, как гравитационные детекторы, в которых используются когерентно осциллирующие твердые — иногда кристаллические — тела[13].)

До сих пор порядки величин выглядят вполне правдоподобно, однако этого, очевидно, недостаточно — необходимо выяснить, выдержит ли идея более суровую проверку. Решающим доказательством могло бы послужить отыскание эксперимен-

тальных ситуаций, в которых возникают, в соответствии с предсказаниями стандартной теории, эффекты, обусловленные макроскопическими квантовыми суперпозициями, но на уровне, на котором, согласно высказанным выше предположениям, такие суперпозиции не могут существовать в течение сколько-нибудь длительного времени. Если в таких ситуациях наблюдение подтвердит традиционные квантовые предположения, то от выдвигаемых мною здесь идей придется отказаться — или, по крайней мере, серьезно их пересмотреть. Если же наблюдение установит, что суперпозиции не сохраняются, то эти идеи получат некоторое достоверное подтверждение. К сожалению, на данный момент я не располагаю сведениями о каких-либо практических предложениях о проведении соответствующих экспериментов. Многообещающие возможности для такого рода экспериментирования предоставляют сверхпроводники и такие устройства, как СКВИДы (сверхпроводящие квантовые интерференционные датчики, в основе действия которых лежат макроскопические квантовые суперпозиции, возникающие в сверхпроводниках); см. [235]. Впрочем, прежде чем приступать непосредственно к экспериментам со сверхпроводниками, предлагаемые идеи следует тщательно доработать. Суперпозиции состояний в сверхпроводнике отличаются очень незначительным смещением масс. Вместо этого здесь имеет место весьма существенное изменение *импульса*, каковая ситуация требует дополнительного теоретического исследования.

Необходимость в некоторой переформулировке вышеизложенной схемы возникает даже в случае простого опыта с камерой Вильсона — иначе, конденсационной камерой, присутствие заряженной частицы в которой сопровождается конденсацией крошечных капель из окружающего частицу пара. Предположим, что заряженная частица находится в квантовом состоянии, представляющем собой линейную суперпозицию состояний «частица находится где-то внутри камеры Вильсона» и «частица находится вне камеры». «Внутренняя» часть вектора состояния частицы инициирует образование капли жидкости, в то время как та часть, согласно которой частица находится снаружи камеры, ничего подобного не делает — т. е. состояние частицы теперь можно рассматривать как суперпозицию двух макроскопически различных состояний. В одном из этих состояний из пара в камере конденсируется капля, в другом — заполняющий камеру пар остается



однородным. Нам же предстоит оценить гравитационную энергию, необходимую для перемещения молекул пара в каждом из образующих суперпозицию состояний. Тут, однако, возникает дополнительное осложнение: следует учесть еще и разницу между *собственной* гравитационной энергией капли и *собственной* гравитационной энергией неконденсированного пара. Для корректного описания таких ситуаций необходима *иная* формулировка предложенного выше критерия. Возможно, здесь следует рассматривать *собственную гравитационную энергию* того распределения масс, которое представляет собой *разницу* между распределениями масс в двух альтернативных состояниях данной квантовой линейной суперпозиции. Таким образом, ожидаемое время редукции будет определяться величиной, обратной этой собственной энергии (см. [300]). В сущности, такая альтернативная формулировка дает в точности тот же результат, что мы уже получили в предыдущих ситуациях, разве что в случае камеры Вильсона время редукции оказывается несколько иным (меньшим). Более того, существуют различные альтернативные общие схемы для определения времени редукции, которые в определенных ситуациях дают различные значения этого самого времени, но которые, тем не менее, вполне согласуются между собой в случае простой суперпозиции двух состояний перемещаемого целиком объекта (см. пример в начале этого параграфа). Первая такая схема была предложена Диози [92] (на некоторые ее недостатки указали Гирарди, Грасси и Римини [147]; они же предложили способ устранения этих недостатков). В последующих главах мы не станем останавливаться на различиях между теми или иными конкретными вариантами, но будем говорить в общем о «предположении (или *критерии*) из § 6.12».

Для чего же нам понадобилось вводить такой особый критерий для «времени редукции»? Мои собственные первоначальные обоснования (см. [295]) носили чересчур специальный характер, чтобы их здесь воспроизводить, и вообще были не очень убедительны и неполны(14). Чуть ниже я приведу независимые аргументы в подтверждение уместности соответствующей физической схемы. Хотя в существующем виде эта аргументация также не совсем полна, она, по всей видимости, все же имеет в своей основе некое мощное требование непротиворечивости, которое дает дополнительное подтверждение предположению о том, что редукция состояний должна, в конечном счете, представлять со-

бой гравитационный феномен, в общем и целом укладывающийся в рамки предлагаемого здесь описания.

О проблеме с *сохранением энергии* в схемах ГРВ-типа мы уже упоминали в § 6.9. «Удары», которым подвергаются частицы (когда их волновые функции самопроизвольно умножаются на гауссову функцию), влекут за собой незначительные нарушения закона сохранения энергии. Более того, передача энергии носит, по всей видимости, *нелокальный* характер. Это, похоже, является характерной — и, вероятно, неизбежной — особенностью общих теорий такого рода, в которых **R**-процедура считается *реальным* физическим эффектом. Мне представляется, что эта особенность может послужить убедительным дополнительным свидетельством в пользу теорий, отводящих ключевую роль в редукции *гравитационным* эффектам, — поскольку в общей теории относительности сохранение энергии всегда было предметом тонким и даже скользким. Гравитационное поле содержит в себе энергию, которая вносит вполне измеримый вклад в общую энергию (и, стало быть, согласно эйнштейновскому $E = mc^2$, массу) системы. С другой стороны, эта энергия представляет собой некую эфемерную субстанцию, существующую в пустом пространстве каким-то загадочным нелокальным образом[15]. Вспомним, в частности, о массе-энергии, что в виде гравитационных волн излучается системой двойного пульсара PSR 1913+16 (см. § 4.5); эти волны суть рябь в самой структуре пустого пространства. Энергия, содержащаяся в полях взаимного притяжения двух нейтронных звезд, также является важной составляющей их динамики, каковую составляющую мы не можем игнорировать. Как раз такая разновидность энергии, «обитающая» в пустом пространстве, и является самой неуловимой из всех. Ее нельзя получить простым «сложением» локальных вкладов плотности энергии, ее даже нельзя локализовать в какой-либо конкретной области пространства-времени (см. НРК, с. 220–221). Возникает искушение соотнести столь же скользкие проблемы нелокальной энергии **R**-процедуры с аналогичными проблемами классической гравитации — сопоставить одни проблемы с другими в надежде разглядеть за ними логически связную общую картину.

Обеспечивают ли такую логическую связность выдвигаемые мною здесь предположения? Думаю, что со временем мы от них этого непременно добьемся, однако на настоящий момент четкой



теоретической основы у нас пока нет. Все, впрочем, говорит за то, что в принципе эта грандиозная задача вполне решаема. В самом деле, как мы уже отмечали ранее, процесс редукции можно сравнить с распадом нестабильной частицы или ядра атома. Представьте себе суперпозицию состояний объекта в двух различных положениях как своего рода нестабильное ядро, распадающееся по истечении некоего характеристического времени «полураспада» на какие-то более стабильные продукты. Аналогичным образом суперпозиция положений объекта — нестабильное квантовое состояние — переходит по истечении некоего характеристического «времени жизни» (определяемого, в грубом приближении, величиной, обратной гравитационной энергии разделения) в состояние стабильное, когда объект оказывается либо в одном положении, либо в другом, что дает нам две возможные формы распада.

Согласно принципу неопределенности Гейзенберга, время жизни (или период полураспада) частицы или ядра атома обратно незначительной *неопределенности* в массе-энергии исходной частицы. (Например, массу нестабильного ядра полония-210, испускающего в процессе распада $\alpha$-частицу и превращающегося в свинец, точно определить невозможно, при этом неопределенность имеет порядок величины, обратной периоду полураспада — в данном случае, около 138 суток, — что дает для полония неопределенность массы всего лишь около $10^{-34}$ общей массы ядра! Для отдельных нестабильных частиц, впрочем, неопределенность составляет существенно бóльшую долю массы.) Таким образом, «распад», сопровождающий процесс редукции, *также* должен предполагать существенную неопределенность энергии исходного состояния. Эта неопределенность, согласно настоящему предположению, обусловлена, по большей части, неопределенностью собственной гравитационной энергии суперпозиции состояний. Собственная же гравитационная энергия включает в себя ту самую эфемерную нелокальную энергию поля, которая уже послужила причиной стольких неприятностей в общей теории относительности и которую нельзя получить простым сложением локальных вкладов плотности энергии. Кроме того, имеется тут и существенная неопределенность в сопоставлении друг другу точек различных пространственно-временны́х геометрий в суперпозиции, что мы отмечали в § 6.10. Если допустить, что *существенная* «неопределенность» энергии состояний в супер-

позиции представлена именно этим гравитационным вкладом, то результат такого допущения вполне согласуется с предсказанным выше временем жизни этого состояния. Таким образом, предлагаемая мною схема позволяет, по всей видимости, убедиться в наличии четкой связи между двумя энергетическими проблемами и по крайней мере обещает возможность построения на основе этих идей вполне непротиворечивой теории.

Наконец, остаются еще два важных вопроса, представляющие для нас в рамках настоящего исследования особый интерес. Первый: каким образом подобные соображения могут помочь нам понять принципы функционирования *мозга*? И второй: есть ли основания (физические) ожидать, что такому гравитационно индуцированному процессу редукции окажется свойственна *невычислимость* (некоего соответствующего вида)? В следующей главе мы увидим, что тут открываются кое-какие весьма захватывающие возможности.

## Примечания

1. Упомянутое в § 5.16 «бозонное» свойство фотонов можно (в некотором смысле) рассматривать как пример проявления квантовой сцепленности, в каковом случае у нас имеется экспериментальное подтверждение и для взаимодействия на сверхбольших расстояниях — результаты наблюдений, полученные Хэнбери Брауном и Твиссом [187, 188] (см. примечание на с. 449).
2. См. [116], [382], [90] и [143].
3. См. [355] и [357].
4. См. [23].
5. В [3] приводится другой весьма серьезный довод в пользу объективной реальности волновой функции.
6. См., например, [82].
7. См. [82], [399], [400] и [283].
8. Именно к этому, похоже, сводятся результаты программы SETI[9], у истоков которой стоял Ф. Дрейк.
9. Мое собственное предположение безоговорочно принадлежит к «гравитационному» лагерю, хотя сколько-нибудь конкретный вид

[9] Search for Extraterrestrial Intelligence — Поиск внеземного разума (англ.) — *Прим. перев.*



оно обрело лишь недавно (см. [295] и [300]). С оригинальным предположением Гирарди — Римини — Вебера его объединяет идея о том, что редукция должна представлять собой внезапный, дискретный процесс. Большинство же современных исследователей, вслед за Перлом [284], склонны рассматривать редукцию состояний как процесс *непрерывный* (стохастический). См. [93], [148] и [303]. Аналогичные рассуждения, но с попыткой сохранения совместимости предлагаемой схемы с теорией относительности, представлены в [149], [151] и [152].

10. [334], также см. НРК, с. 290–296.
11. См. также [92], [147] и [295].
12. См. [392].
13. См. [379], [39].
14. Впрочем, похоже, что предложенный здесь критерий отвечает общим требованиям, изложенным в НРК (глава 7), гораздо лучше (как я, собственно, и предполагал в [295]), нежели сформулированный все в том же НРК «одногравитонный критерий». Для того, чтобы составить об этом соответствии более конкретное представление, необходимы дополнительные исследования.
15. См. [293]; а также НРК, с. 220–221.

# 7

# КВАНТОВАЯ ТЕОРИЯ И МОЗГ

## 7.1. Макроскопическая квантовая процедура в работе мозга

Согласно общепринятой точке зрения, понимание (истинное или кажущееся) работы мозга следует искать в рамках классической физики. Считается, что передаваемые по нервам сигналы суть феномены типа «есть или нет», точно так же, как токи в электронных цепях компьютера — они *либо* есть, *либо* их нет, здесь не бывает тех таинственных *суперпозиций* альтернативных вариантов, что характерны для квантовой физики. Хотя на фундаментальном уровне квантовые эффекты, вероятно, играют определенную роль, биологи в большинстве своем придерживаются мнения, что при рассмотрении макроскопических следствий примитивных квантовых закономерностей необходимости выходить за классические рамки нет. Химические силы, управляющие межатомными и межмолекулярными взаимодействиями, и впрямь имеют квантовомеханическое происхождение, и именно химические взаимодействия определяют по большей части поведение *нейромедиаторов*, передающих сигналы от одного нейрона к другому через узкие промежутки между ними (так называемые *синаптические щели*). Аналогичным образом, потенциалы действия, физически контролирующие передачу нервных импульсов, имеют предположительно квантовомеханическую природу. И все же мы, как правило, допускаем, что и поведение отдельных нейронов, и их взаимодействие вполне адекватно моделируются классическим средствами. Соответственно, широко распространено мнение, что модель физической деятельности мозга как целого следует строить по *классическим* «правилам», не обращая



особого внимания на тонкие и загадочные эффекты квантовой физики.

Отсюда непосредственно следует, что с точки зрения наблюдателя любой существенный процесс в мозге *либо* «происходит», *либо* «не происходит». Странные *суперпозиции* квантовой теории, допускающие ситуации, когда процесс *одновременно* «происходит» *и* «не происходит», — и снабженные соответствующими комплексными весовыми коэффициентами — естественно, в расчет не принимаются. Мы еще можем согласиться с тем, что на некоем субмикроскопическом уровне подобные квантовые суперпозиции «действительно» имеют место, однако на уровне макроскопическом, по нашему глубокому убеждению, характерные для таких квантовых феноменов эффекты интерференции сколько-нибудь существенной роли играть просто не могут. Следовательно, любые такие суперпозиции уместно рассматривать как статистические эффекты, а классическое моделирование функционирования мозга оказывается с практической точки зрения (и снова FAPP!) целиком и полностью удовлетворительным.

Однако такого мнения придерживаются далеко не все. В частности, известный нейрофизиолог Джон Эккльз указывал на важную роль квантовых эффектов в синаптической передаче (см., например, [18] и [105]). По предположению Эккльза, квантовая активность сосредоточена в так называемой пресинаптической везикулярной сетке — паракристаллической гексагональной структуре в пирамидальных клетках мозга. Другие ученые (включая и меня, см. НРК, с. 400–401 и [291]), экстраполируя тот факт, что светочувствительные клетки сетчатки (которая формально является частью мозга) способны реагировать на чрезвычайно слабый свет (буквально несколько фотонов, [194]) — при определенных обстоятельствах такая клетка может зарегистрировать даже *один-единственный* фотон [17], — предположили, что и в самом мозге могут содержаться нейроны, также являющиеся, по сути своей, квантовыми «детекторами».

Поскольку квантовые эффекты действительно могут инициировать в мозге процессы гораздо более крупного, нежели сами, «масштаба», отдельные исследователи выразили надежду, что способность *разума* воздействовать на физический мозг может быть обусловлена *квантовой неопределенностью*. Здесь сле-

дует, скорее всего, принять — явно или нет — *дуалистическую* точку зрения. Вполне возможно, что на квантовые вероятности, реально возникающие в результате таких недетерминированных процессов, оказывает влияние «свободная воля» «внешнего разума». В этом случае, «материя разума» нашего дуалиста воздействует на поведение его физического мозга не иначе, как через посредство квантовой **R**-процедуры.

Я не знаю, как относиться к подобным предположениям, особенно в свете того, что в стандартной квантовой теории никакой неопределенности на квантовом уровне *нет* — здесь действует вполне детерминированная **U**-эволюция. Предполагается, что неопределенность, связанная с процедурой **R**, возникает лишь в процессе перехода с квантового уровня на классический. Согласно стандартному FAPP-объяснению, неопределенность эта «происходит» лишь тогда, когда квантовое событие оказывается сцепленным с достаточным объемом окружения. Более того, как мы могли убедиться в § 6.6, само понятие «происходить» трактуется в стандартном подходе крайне туманно. Вряд ли в рамках традиционной квантовой физики можно утверждать, что теория допускает-таки существование неопределенности на уровне единичной квантовой частицы — такой, например, как фотон, атом или небольшая молекула. Например, встреча волновой функции фотона с фоточувствительной ячейкой инициирует целую последовательность событий, которые остаются детерминированными (эволюция **U**), пока система пребывает «на квантовом уровне». Затем возмущение охватывает достаточный объем окружения, и мы говорим, что произошла (FAPP) редукция **R**. Придется смириться с тем, что «материя разума» способна так или иначе воздействовать на систему лишь на этой стадии неопределенности.

Согласно моему собственному представлению о редукции состояний (см. § 6.12), в поисках уровня, на котором действительно происходит **R**-процесс, следует обратить внимание на масштабы вполне макроскопические, что имеет смысл, когда в квантовом состоянии оказываются сцепленными довольно большие объемы вещества (от нескольких микрон до нескольких миллиметров в диаметре — или даже гораздо бóльшие, если процесс не предполагает значительного перемещения масс). (В дальнейшем я буду называть эту вполне конкретную, но, тем не менее, гипотетическую «действующую» редукцию *объективной* и



обозначать через **OR**[1].) В любом случае, если мы собираемся придерживаться описанной выше дуалистической точки зрения, где нам нужно еще отыскать «место», откуда внешний «разум» сможет воздействовать на физическое поведение мозга, — для успешного поиска придется, по-видимому, заменить чистую случайность квантовой теории чем-то более утонченным, — то мы непременно должны выяснить, каким образом воздействие «разума» может проявляться в масштабах, существенно более крупных, нежели размер отдельной квантовой частицы. Искать ответ следует там, где квантовый и классический уровни соприкасаются. Трудность заключается в том, что мы, как уже отмечалось в предыдущей главе, никак не можем договориться о том, существует ли такая точка соприкосновения вообще, а если существует, то что она собой представляет и где находится.

Думаю, что с научной точки зрения довольно бессмысленно полагать, что дуалистический «разум», *внешний* (что логично) по отношению к телу, каким-то загадочным образом воздействует на выбор того или иного альтернативного варианта, происходящий, судя по всему, под действием процедуры **R**. Если бы «воля» могла каким-то образом изменять выбор, который осуществляет в момент **R** Природа, то почему же экспериментатор не может с помощью своей «силы воли» воздействовать на результат квантового эксперимента? Если бы такое было возможно, то нарушения квантовой вероятности происходили бы сплошь и рядом! Лично я, как ни пытаюсь, не могу поверить в то, что подобная картина может быть хоть сколько-нибудь близка к реальности. Представление о внешней «материи разума», не подвластной физическим законам, выводит нас за рамки того, что можно обоснованно назвать научным объяснением, отсылая прямиком к точке зрения $\mathscr{D}$ (см. § 1.3).

[1] В НРК я использовал для обозначения такого процесса термин «корректная квантовая гравитация» (ККГ[2]). Здесь же акцент несколько иной. Сейчас я не хочу указывать на связь рассматриваемой процедуры с фундаментальной задачей построения непротиворечивой теории квантовой гравитации. Я хочу, скорее, подчеркнуть, что в основе этой процедуры лежат те же предположения, что я сделал в § 6.12, плюс некий фундаментальный неизвестный и невычислимый компонент. Использование сокращения **OR**[3] имеет еще и дополнительный смысл: физическим результатом объективной редукции и в самом деле является *одно* состояние — *или* то, *или* другое, — в отличие от комплексной суперпозиции, с которой мы имели дело прежде.

[2] Англ. CQG, *correct quantum gravity*. — *Прим. перев.*

[3] Англ. *or* переводится как «или». — *Прим. перев.*

Впрочем, однозначно оспорить такую точку зрения очень сложно, так как по самой своей природе она лишена четких правил, которые позволили бы нам подойти к ней с позиций строгого научного рассуждения. Тех читателей, которые по каким-либо причинам твердо убеждены, что наука никогда не дорастет до того, чтобы хотя бы подступиться к проблемам разума (точка зрения $\mathscr{D}$), я смиренно прошу потерпеть меня еще немного и просто посмотреть, какие «пустоты» могут в самое ближайшее время обнаружиться в монолите современной науки и, несомненно, послужить ее распространению далеко за пределы тех тесных границ, которые она на сегодняшний день для себя установила. Если «разум» представляет собой нечто внешнее по отношению к физическому телу, то почему же тогда столь многие его качества так тесно связаны со свойствами физического мозга? Моя собственная точка зрения заключается в том, что для отыскания ответа на этот и другие подобные вопросы необходимо более тщательно исследовать известные физические «материальные» структуры, составляющие мозг, — и разобраться, наконец, *что* же в действительности *представляют собой* «материальные» структуры на квантовом уровне. Полагаю, иного выхода у нас, в конечном счете, нет — чтобы добраться до истины, нам придется углубиться в самые основы мироздания.

Как бы то ни было, ясно по крайней мере одно. Мы должны рассматривать не просто квантовые свойства отдельных частиц, атомов или даже малых молекул, но эффекты квантовых систем, сохраняющие свою явно квантовую природу на макроскопическом уровне. Если в системе отсутствует макроскопическая квантовая когерентность, то неоткуда взяться и тонким эффектам на квантовом уровне — таким, скажем, как нелокальность и квантовый параллелизм (несколько одновременных действий в суперпозиции), — или эффектам контрфактуальности, приобретающих значимость лишь на классическом уровне функционирования мозга. Без должного «экранирования» квантового состояния от окружения такие эффекты мгновенно затеряются в присущей этому окружению хаотичности, — выражающейся, в нашем случае, в беспорядочном движении молекул биологических веществ и жидкостей, составляющих основную массу мозга.

Что же такое *квантовая когерентность*? Этот феномен возникает при условиях, позволяющих большому количеству частиц образовывать совместно единое квантовое состояние, практически несцепленное с окружением. (Термином «когерентность» в общем случае обозначается согласованность отдельных колебаний по фазе. Говоря о *квантовой* когерентности, мы имеем в виду колебательную природу волновой функции; когерентность в данном случае подразумевает наличие единого квантового состояния.) Такие состояния в наиболее наглядном виде встречаются в феноменах сверхпроводимости (когда электрическое сопротивление проводника равно нулю) и сверхтекучести (когда равно нулю жидкостное трение, или вязкость). Характерной особенностью таких феноменов является наличие *запрещенной энергетической зоны* — для того чтобы изменить существующее квантовое состояние, окружение должно эту зону как-то преодолеть. Когда температура окружения достаточно высока, т. е. частицы, это окружение составляющие, обладают энергией, достаточной для того, чтобы «перепрыгнуть» запрещенную зону и «сцепиться» с квантовым состоянием, квантовая когерентность разрушается. Поэтому явления, подобные сверхпроводимости и сверхтекучести, возникают обычно лишь при очень низких температурах, порядка нескольких градусов выше абсолютного нуля. В этом, собственно, и заключалась (до недавних пор) одна из причин общего скептического отношения к возможности существования эффектов квантовой когерентности внутри такого «горячего» объекта, как человеческий мозг — или любая другая биологическая система.



Однако за последние годы было проведено несколько замечательных экспериментов, показавших, что в некоторых веществах сверхпроводимость может возникать при гораздо более высоких температурах, вплоть до 115 К (см. [343]). С биологической точки зрения, это все еще слишком холодно: −158°C (или −212°F) — лишь немногим выше температуры жидкого азота. Гораздо более интересны в этом смысле наблюдения Лаге и его коллег [233], указывающие на существование сверхпроводимости при температурах всего лишь «сибирских», −23°C (или −10°F).

Будучи все еще несколько, по биологическим меркам, «холодноватой», такая *высокотемпературная сверхпроводимость* является серьезным свидетельством в пользу предположения о возможности существования квантовокогерентных эффектов в биологических системах.

Более того, еще задолго до обнаружения феномена высоко-

температурной сверхпроводимости выдающийся физик Герберт Фрёлих (совершивший в 1930-е годы один из фундаментальных «прорывов» в понимании «обычной» низкотемпературной сверхпроводимости) предположил, что коллективные квантовые эффекты могут играть определенную роль в биологических системах. Заинтересовавшись необычным феноменом, наблюдавшимся еще в 1938 году на биологических мембранах (и применив концепцию, предложенную Ларсом Онсагером и моим братом, Оливером Пенроузом [289], — о чем я, занявшись изучением вопроса, узнал с некоторым удивлением), Фрёлих в 1968 году [129] пришел к выводу, что биологическая квантовая когерентность должна вызывать в живых клетках колебательные эффекты, резонирующие с микроволновым электромагнитным излучением на частоте $10^{11}$ Гц. Эти эффекты не требуют низких температур и возникают благодаря большой энергии метаболических процессов. Сегодня мы располагаем достоверными экспериментальными свидетельствами, подтверждающими наличие во многих биологических системах в точности таких эффектов, какие предсказывал в 1968 году Фрёлих. Чуть позже (в § 7.5) мы попробуем разобраться, какое отношение эти феномены могут иметь к работе мозга.

## 7.2. Нейроны, синапсы и компьютеры

Получить явное подтверждение тому, что квантовая когерентность действительно может играть в биологических системах ключевую роль, конечно же, отрадно, однако суть этой самой роли применительно к процессам, имеющим непосредственное отношение к функционированию мозга, пока совершенно не ясна. Наше понимание работы мозга, все еще очень смутное, сводится, по большей части, к классическому представлению (совпадающему, в основном, с тем, что предложили еще в 1943 году Маккаллох и Питтс), согласно которому нейроны и соединяющие их синапсы выполняют в мозге практически те же функции, что и транзисторы вместе с соединяющими их дорожками в печатных схемах современных компьютеров. Более детальная биологическая картина выглядит так: классические нервные сигналы распространяются из центрального тела нейрона (*сомы*) вдоль очень длинного волокна, называемого *аксоном*, причем от аксона в различных местах ответвляются отдельные отростки (см. рис. 7.1). Каждый отросток непременно заканчивается *синапсом* — соединением, посредством которого сигнал через синаптическую щель передается к следующему нейрону (как правило). Именно на этой стадии в процесс вступают химические вещества, называемые *нейромедиаторами*, — перемещаясь от одной клетки (нейрона) к другой, они переносят сообщение о возбуждении предыдущего нейрона. Такое синаптическое соединение приходится либо на древовидный отросток (*дендрит*) следующего нейрона (в большинстве случаев), либо на его сому. Одни синапсы являются по своей природе возбуждающими, их нейромедиаторы усиливают возбуждение следующего нейрона; другие же, напротив, — тормозящие, и их нейромедиаторы (отличные от первых) возбуждение следующего нейрона ослабляют. Воздействие различных синапсов на нейрон суммируется (возбуждение учитываем со знаком «плюс», а торможение — со знаком «минус»), и по достижении определенного порогового значения нейрон возбуждается[4]. Правильнее, впрочем, будет сказать, что существует высокая *вероятность* такого возбуждения. Определенный случайный фактор присутствует во всех процессах такого рода.



Таким образом — во всяком случае, пока, — не возникает сомнений в том, что изложенная картина может быть эффективно смоделирована численными методами, если допустить, что синаптические связи и их индивидуальная интенсивность со временем не изменяются. (Наличие случайных составляющих, разумеется, никаких проблем в смысле вычислимости не представляет, см. § 1.9). В самом деле, несложно заметить, что вышеописанная нейронно-синапсовая схема (с постоянными синапсами и их интенсивностями) существенно *эквивалентна* схеме компьютера (см. НРК, с. 392–396). Однако благодаря феномену так называемой *пластичности мозга*, интенсивность по крайней мере

[4] По крайней мере, таково традиционное представление. Сегодня у нас есть некоторые основания полагать, что эта простая «аддитивная» модель слишком упрощена и определенная «обработка информации» может осуществляться уже в дендритах отдельных нейронов. На возможность такой обработки указывал, среди прочих, Карл Прибрам (см. [319]). Сходные в общих чертах предположения были сделаны ранее Алвином Скоттом [338, 339] (а о возможности наличия «интеллекта» в отдельно взятой клетке можно прочесть, например, у Альбрехта-Бюлера [8]). Возможность сложной «дендритной» обработки информации внутри отдельных нейронов мы подробнее обсудим в § 7.4.

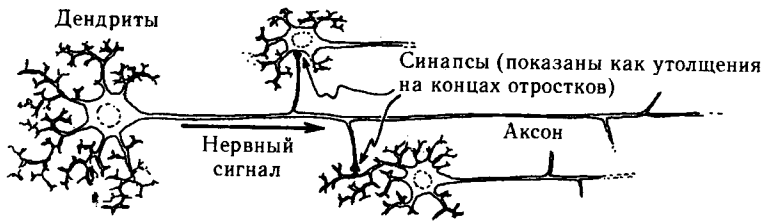



Рис. 7.1. Нейрон и его соединение с другими нейронами посредством синапсов.

некоторых синаптических связей может время от времени изменяться — порой быстрее, чем за секунду, — а кроме того, изменяться могут и сами связи. Что ставит нас перед немаловажным вопросом: что же этими синаптическими изменениями управляет?

В коннекционистских моделях (применяемых при разработке искусственных нейронных сетей) синаптические изменения описываются определенным *вычислительным правилом*. Это правило устанавливается таким образом, чтобы система могла в процессе работы повышать свою эффективность, сравнивая поступающую на ее вход извне информацию с некоторыми заранее заданными критериями. Простое правило такого типа предложил Дональд Хебб еще в 1949 году [193]. Современные коннекционистские модели(1) используют различные модификации (порой весьма значительные) все той же процедуры Хебба. Любая модель такого рода непременно должна иметь в своей основе *хоть какое-нибудь* четкое вычислительное правило, поскольку выполняются эти модели на самых обычных компьютерах; см. § 1.5. Однако, в силу изложенной в первой части аргументации, никакая вычислительная процедура не может адекватно объяснить все операционные проявления человеческого сознательного понимания. Следовательно, нужно искать какой-то другой управляющий «механизм» — по крайней мере, для объяснения синаптических изменений, возможно, имеющих некоторое отношение к настоящей *сознательной* деятельности мозга.

Были выдвинуты и другие идеи; например, Джеральд Эдельман в своей книге «Прозрачный воздух, сверкающий огонь» [112] (и в более ранней трилогии [109, 110, 111]) предположил, что в мозге действуют не правила типа правила Хебба, а, скорее, некий



вариант «дарвиновского» эволюционного принципа, позволяющий мозгу непрерывно повышать свою эффективность, управляя синаптическими связями посредством своеобразного естественного отбора, — при этом Эдельман указывает на весьма многозначительные параллели между своей моделью и процессом развития иммунной системой способности «распознавать» вещества. Особое значение в этой модели придается сложной роли нейромедиаторов и других химических соединений, задействованных в коммуникации между нейронами. Однако на сегодняшний день соответствующие процессы по-прежнему рассматриваются как классические и вычислимые. Вместе со своими коллегами Эдельман даже построил ряд устройств с компьютерным управлением (получивших названия DARWIN I, II, III, IV и т. д.), предназначенных для моделирования (с увеличением степени сложности) как раз той самой процедуры, которая, по его предположению, лежит в основе умственной деятельности. Однако тот факт, что управляющие функции в устройствах Эдельмана возложены на самый обычный универсальный компьютер, вполне недвусмысленно показывает, что и эта схема является исключительно вычислительной — просто здесь используется некая «восходящая» система правил. При этом совершенно не важно, какими именно деталями данная схема отличается от других вычислительных процедур. Она все равно принадлежит к той категории, что мы обсуждали в первой части, — см. § 1.5, а также § 3.9 и краткое изложение аргументации главы 3 в воображаемом диалоге в § 3.23. Одного лишь этого диалога достаточно для того, чтобы убедиться в полном неправдоподобии любого утверждения о том, что модель, основанная только на подобного рода принципах, может иметь какое-то отношение к действительному функционированию сознательного разума.

Для того, чтобы избавиться от этих «пут» вычислительности, необходимо найти какой-нибудь другой механизм управления синаптическими связями — причем каким бы этот механизм ни был, он, по всей видимости, должен задействовать некий физический процесс, важную роль в котором играет та или иная форма квантовой когерентности. Если этот процесс окажется в каком-либо существенном отношении похожим на действие иммунной системы, то, значит, и иммунная система работает на квантовых эффектах. Возможно, какие-то процессы в работе иммунного механизма распознавания и впрямь носят существенно квантовый ха-

рактер — как, в частности, утверждает Майкл Конрад [57, 58, 59]. Меня бы это не удивило, однако в эдельмановской модели мозга возможному участию квантовых процессов в работе иммунной системы места не нашлось.

Впрочем, даже если когерентные квантовомеханические эффекты каким-то образом замешаны в управлении синаптическими связями, все же трудно предположить, что и распространение нервных импульсов может быть связано с чем-то существенно квантовомеханическим. Иначе говоря, совершенно неясно, какую пользу можно извлечь из рассмотрения квантовой суперпозиции, в которой нейрон одновременно и *возбужден*, и *заторможен*. Нервные сигналы представляются нам явлениями вполне макроскопическими — во всяком случае, достаточно макроскопическими для того, чтобы такая картина выглядела крайне неправдоподобно, даже несмотря на тот факт, что собственно передача весьма хорошо изолирована от окружения благодаря плотному слою миелина, покрывающему нервные окончания. Согласно критерию, предложенному в §6.12 (**OR**), следует ожидать, что при возбуждении нейрона объективная редукция состояния происходит очень быстро — не потому, что имеет место значительное перемещение масс (его там даже по минимально требуемым стандартам далеко недостаточно), а потому, что распространяющееся вдоль нерва электрическое поле (порождаемое нервным сигналом), скорее всего, не остается «незамеченным» окружающими нерв тканями мозга. Это поле возмущает случайным образом весьма значительный объем вещества окружения — вполне достаточный, как мне представляется, для того, чтобы удовлетворить критерию срабатывания процедуры **OR** (из §6.12) почти сразу же после возникновения сигнала. Таким образом, сохранение в течение длительного времени квантовых суперпозиций возбуждения и торможения нейрона вряд ли возможно.

## 7.3. Квантовые вычисления

Свойство возбужденного нейрона возмущать окружение всегда представлялось мне донельзя неудобным — оно никак не вписывалось в то предварительное предположение, которое я пытался обосновать в НРК и в рамках которого квантовая суперпозиция одновременного возбуждения и торможения семейств нейронов была, как мне казалось, действительно необходимой. Согласно нашему новому критерию редукции состояний (**OR**),



для редукции требуется еще меньшее возмущение окружения, чем в прежнем описании, и в возможность сохранения таких суперпозиций в течение сколько-нибудь заметного времени поверить еще сложнее. А собственно идея тогда заключалась в следующем: если бы возможно было выполнять несколько отдельных «вычислений» в суперпозиции в нескольких одновременно возбуждающихся нейронных структурах, то резонно было бы предположить, что в мозге вместо «обычных» тьюринговых вычислений выполняется нечто вроде вычислений *квантовых*. Несмотря на кажущуюся невозможность выполнения квантовых вычислений на этом уровне функционирования мозга, будет полезно познакомиться с некоторыми их аспектами подробнее.

Квантовое вычисление — теоретическая концепция, основы которой разработали Дэвид Дойч [83] и Ричард Фейнман [120, 121] (см. также [25] и [6]) и которая в настоящее время активно исследуется многими учеными. Основная идея заключается в распространении классического понятия машины Тьюринга на соответствующее квантовое устройство. Как следствие, все выполняемые такой расширенной «машиной» операции должны подчиняться квантовым законам — т. е. законам, по которым живут системы квантового уровня (с возможностью суперпозиций). Так, эволюция устройства происходит преимущественно под действием процедуры **U**, причем существенным свойством этого самого действия является как раз сохранение наличествующих суперпозиций. Процедура **R** получает «право голоса», как правило, лишь в *конце* операции, когда система «измеряется» с целью узнать результат вычисления. Вообще говоря (хотя не все это осознают), в процессе вычисления процедуру **R** необходимо время от времени вызывать дополнительно для того, чтобы проверить, не завершилось ли оно.

Выяснилось, что, хотя квантовый компьютер и не имеет сверхспособностей, *в принципе* недоступных для традиционного вычисления по Тьюрингу, в некоторых классах задач квантовое вычисление превосходит тьюрингово вычисление в смысле *теории сложности* ([83]). То есть при решении таких задач квантовый компьютер оказывается в принципе намного *быстрее*, нежели компьютер обычный, — *но и только*. Ряд интересных (хотя и несколько искусственных) задач такого типа, при решении которых квантовый компьютер оказывается победителем, приводят, в частности, Дойч и Йожа [88]. Более того, как недавно

показал Питер Шор, с помощью квантового вычисления можно решить (за полиномиальное время) актуальную задачу факторизации больших целых чисел.

«Стандартное» квантовое вычисление использует обычные правила квантовой теории, согласно которым в течение практически всей операции система эволюционирует под действием процедуры **U**, а **R** вмешивается в процесс на строго определенных этапах. В такой процедуре нет ничего «невычислимого» в смысле *обычной* «вычислимости», так как **U** — вычислимая операция, а **R** — чисто вероятностная процедура. Все, что в принципе можно получить с помощью квантового компьютера, можно в принципе получить и с помощью соответствующей машины Тьюринга, снабженной генератором случайных чисел. Таким образом, согласно представленным в первой части книги аргументам, даже квантовый компьютер не способен выполнять операции, требуемые для человеческого сознательного понимания. Остается надеяться лишь на то, что *подлинная* невычислимость скрывается где-то за тонкими особенностями процесса, *в действительности* происходящего в момент «кажущейся» редукции вектора состояния, потому что во временно заменяющей этот реальный процесс случайной процедуре **R** никакой невычислимости нет. Таким образом, полная теория гипотетической процедуры **OR** будет по необходимости носить *существенно невычислимый* характер.

Предложенная в НРК идея основывалась на предположении, что в мозге возможны достаточно длительные тьюринговы вычисления в суперпозиции, прерываемые время от времени неким невычислимым действием, которое можно объяснить лишь в терминах того нового физического процесса (например, **OR**), какой придет на смену редукции **R**. Теперь, когда на такие суперпозиции нейронных вычислений мы больше рассчитывать не можем по причине слишком сильного возмущения окружения проходящими по нейрону импульсами, становится непонятно, каким образом можно здесь хотя бы воспользоваться самой идеей стандартного квантового вычисления, не говоря уже о какой-либо модификации этой процедуры посредством замены **R** на некий гипотетический невычислимый процесс (например, **OR**). Однако, как мы очень скоро убедимся, существует еще одна, весьма многообещающая возможность. Для того чтобы понять, что она собой представляет, нам необходимо более подробно рассмотреть биологическое устройство клеток мозга.



## 7.4. Цитоскелет и микротрубочки

Если мы вдруг вообразим, что сложное поведение животных управляется только лишь нейронами, то скромная парамеция поставит нас перед фундаментальной проблемой. Эта инфузория перемещается по своему пруду с помощью многочисленных крохотных волосообразных конечностей — *ресничек*, — преследуя бактерий, которыми она питается и которых обнаруживает посредством различных внутренних механизмов, или отступая от возможной опасности, готовая мгновенно устремиться прочь. Она также может преодолевать препятствия, огибая их. Более того, парамеция, по всей видимости, способна обучаться на собственном опыте[2] — хотя эта наиболее замечательная ее способность некоторыми учеными оспаривается[3]. Как же все это может проделывать существо, не имеющее ни единого нейрона и синапса? В самом деле, поскольку вся парамеция — это всего лишь одна, пусть и большая, клетка, и притом не нейрон, ей просто негде все перечисленные способности разместить (см. рис. 7.2).

Несомненно, поведение парамеции — да собственно и прочих одноклеточных организмов, например, амеб — регулируется какой-то сложной системой управления, просто эта система построена не из нервных клеток. Ответственная за поведение парамеции структура, очевидно, является частью ее так называемого *цитоскелета*. Как можно предположить из названия, цитоскелет служит для поддержания формы клетки, однако у него имеются и многочисленные иные функции. Упоминавшиеся выше реснички представляют собой окончания волокон цитоскелета, но помимо них цитоскелет, похоже, содержит еще и собственно систему управления движением клетки, а также систему «конвейеров», осуществляющих транспортировку молекул внутри клетки. Словом, в единичной клетке цитоскелет выступает в роли этакой комбинации скелета, мускулатуры, конечностей, системы кровообращения и нервной системы.

Нас с вами в настоящий момент больше всего интересует, каким образом цитоскелет выполняет функции клеточной «нервной системы». Нейроны в нашем мозге сами являются отдельными клетками, причем у каждого нейрона есть свой *собственный* цитоскелет! Означает ли это, что в некотором смысле каждый отдельный нейрон располагает чем-то вроде «личной нервной си-

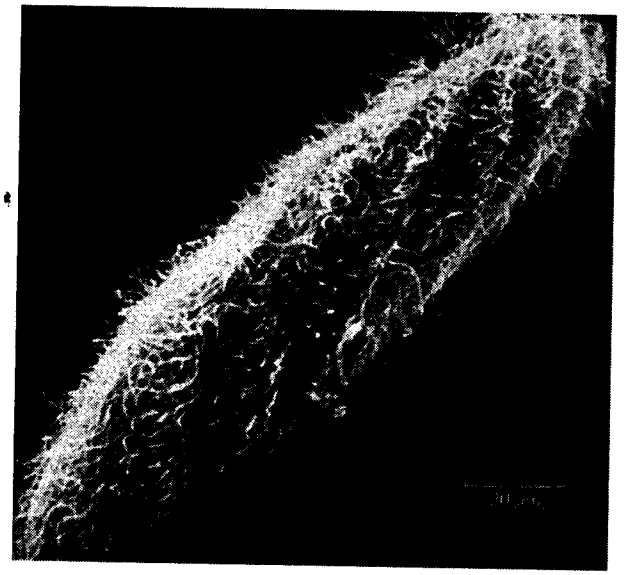


Рис. 7.2. *Парамеция.* Обратите внимание на волосообразные реснички, используемые для перемещения в воде. Они представляют собой наружные окончания *цитоскелета* парамеции.

стемы»? Предположение весьма интригующее, и многие ученые склоняются к мнению, что нечто подобное действительно может иметь место. (См. первопроходческий труд Стюарта Хамероффа «Первичное вычисление: биомолекулярное сознание и нанотехнология» [183]; также рекомендую обратить внимание на статью [184] и многочисленные статьи в новом журнале «Нанобиология»[5].)

Прежде чем переходить к этим вопросам, необходимо рассмотреть вкратце общее устройство цитоскелета. Он состоит из протеиноподобных молекул, организованных в различного типа структуры: актин, микротрубочки и промежуточные волокна. Нас сейчас интересуют, главным образом, *микротрубочки*. Они представляют собой полые цилиндрические трубки с внешним

[5] *Nanobiology.*



диаметром около 25 нм и внутренним — около 14 нм (где «нм» обозначает «нанометр», т. е. $10^{-9}$ м), иногда организованные в более крупные трубкообразные волокна, состоящие из девяти дублетов, триплетов или частичных триплетов микротрубочек; в поперечном сечении такое волокно напоминает лопасти вентилятора, как показано на рис. 7.3, причем иногда по его центру также проходит пара микротрубочек. Как раз такое строение имеют реснички парамеции. Каждая микротрубочка представляет собой белковый полимер, состоящий из субъединиц, называемых *тубулинами*. Каждая субъединица тубулина, в свою очередь, представляет собой «димер», т. е. состоит из двух соединенных тонкой перемычкой частей, называемых $\alpha$-тубулин и $\beta$-тубулин (приблизительно по 450 аминокислот в каждой). Эти, пары глобулярных белков, напоминающие по форме орех арахиса, уложены в слегка скошенную гексагональную решетку вдоль всей трубки, как показано на рис. 7.4. Обычно на каждую миктротрубочку приходится по 13 рядов димеров тубулина. Размеры димера составляют приблизительно 8 нм $\times$ 4 нм $\times$ 4 нм, а его атомное число — около $11 \times 10^4$ (т. е. в одном димере содержится такое количество нуклонов, что его масса в абсолютных единицах равна приблизительно $10^{-14}$).

Димер тубулина может существовать в двух (по крайней мере) различных геометрических конфигурациях, называемых *конформациями*. В одной из таких конформаций молекулы тубулина располагаются под углом около 30° к оси микротрубочки. Есть основания полагать, что эти две конформации соответствуют двум различным состояниям электрической поляризации димера, возникающим вследствие того, что электрон в центре перемычки $\alpha$-тубулин/$\beta$-тубулин занимает в различных конформациях различные положения.

«Центром управления» в цитоскелете является, по всей видимости, структура, называемая *центром организации микротрубочек*, или *центросомой*. Внутри центросомы имеется особая структура, называемая *центриолью*, которая состоит из двух цилиндрических волокон, по девять триплетов микротрубочек в каждом, образующих в пространстве структуру, похожую на «разделенную» букву «Т» (см. рис. 7.5). (Цилиндрические волокна в общем аналогичны по структуре ресничкам, показанным на рис. 7.3.) Согласно Альбрехту-Бюлеру [7, 9], центриоль действует как глаз (!) клетки — идея чрезвычайно захватывающая, хотя и

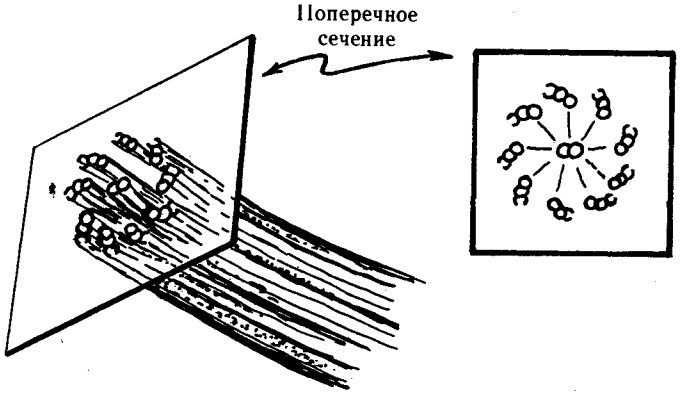



Рис. 7.3. Важной частью цитоскелета являются пучки крохотных трубочек (микротрубочек), организованных в структуры, напоминающие в поперечном сечении лопасти вентилятора. Такое строение имеют, например, реснички парамеции.

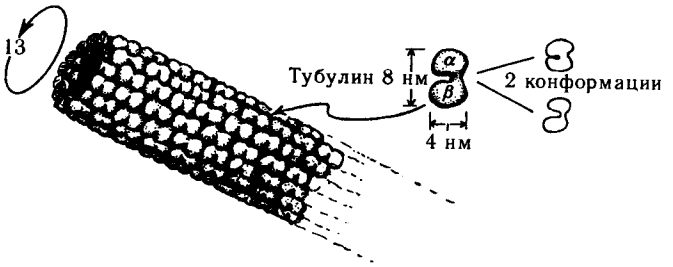



Рис. 7.4. *Микротрубочка*. Полая трубка, обычно состоящая из 13 рядов димеров тубулина. Каждая из молекул тубулина может существовать в двух (по крайней мере) конформациях.

далеко еще не общепринятая. Какой бы ни была роль центросомы в нормальной, «повседневной», жизни клетки, она выполняет по крайней мере одну фундаментально важную задачу. На некоем критическом этапе она разделяется на две части, каждая из которых, по всей видимости, утягивает за собой пучок микротрубочек — хотя, пожалуй, точнее будет сказать, что каждая часть становится своего рода фокусом, вокруг которого и собираются микротрубочки. Эти микротрубочковые волокна каким-то образом связывают центросому с отдельными цепочками ДНК в ядре (в центральных точках, называемых центромерами), и цепочки ДНК расходятся — начиная тем самым удивительный процесс, известный специалистам под названием *митоз*, что означает всего-навсего *деление клетки* (см. рис. 7.6).



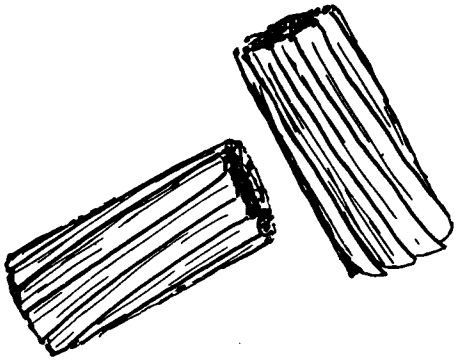


Рис. 7.5. *Центриоль* (по некоторым предположениям, *глаз* клетки) состоит из двух пучков микротрубочек (очень похожих на те, что изображены на рис. 7.3), образующих «разделенную» букву «Т».

Может показаться странным, что внутри одной клетки действуют две столь разные «штаб-квартиры». Одна из них — *ядро*, где хранится основной генетический материал клетки, определяющий ее наследственность и уникальность, а также управляющий производством белкового материала, из которого, собственно, «строится» клетка. Другой управляющий центр — *центросома* с *центриолью* в качестве основного компонента, являющаяся, по всей видимости, главным узлом цитоскелета — структуры, которая, опять же по всей видимости, контролирует движение клетки и ее пространственную организацию. Предполагается, что присутствие этих двух различных «центров» в эукариотических клетках (клетках всех животных и почти всех

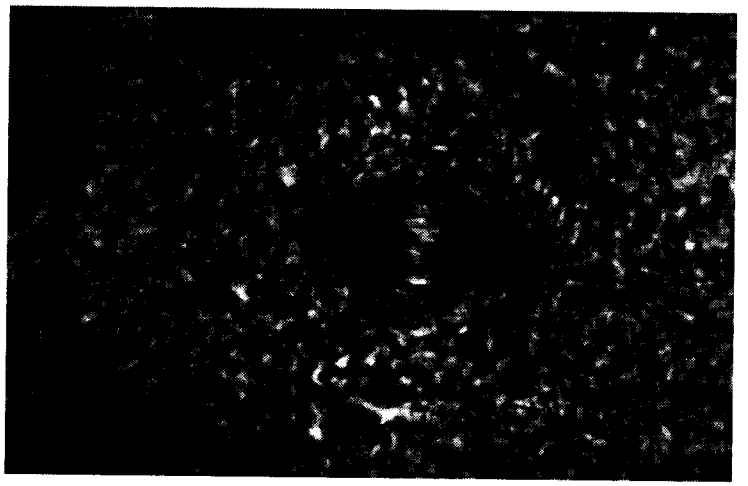


Рис. 7.6. При митозе (делении клетки) хромосомы разделяются, растаскиваемые пучками микротрубочек

растений на нашей планете, за исключением бактерий, сине-зеленых водорослей и вирусов) является результатом древней «инфекции», распространившейся по миру несколько миллиардов лет назад. Клетки, населявшие Землю прежде, были прокариотическими; они существуют и поныне в виде бактерий и сине-зеленых водорослей, и у них нет цитоскелета. Согласно одному из предположений [332], часть древнейших прокариот оказались каким-то образом связаны (возможно, «инфицированы») с неким видом спирохет (бактерий, перемещающихся с помощью нитеобразного хвоста, состоящего из цитоскелетных белков). Эти чуждые друг другу организмы постепенно «научились» жить вместе в симбиотической связи как единые *эукариотические* клетки. Так «спирохеты» превратились, в конечном счете, в цитоскелеты клеток — со всеми вытекающими последствиями для будущей эволюции, среди которых мы с вами!

Организация микротрубочек млекопитающих представляет интерес с математической точки зрения. На первый взгляд, число 13 не имеет какого-либо особого математического значения, однако это не совсем так. Оно принадлежит к знаменитой последовательности *чисел Фибоначчи*:



$$0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, \ldots$$

где каждое последующее число получается сложением двух предыдущих. Это может показаться случайным совпадением, однако хорошо известно, что числа Фибоначчи в биологических системах не редкость (и в гораздо более крупном масштабе). Например, в еловых шишках, цветках подсолнечника и пальмовых стволах наблюдаются спиральные или винтовые структуры с взаимопроникновением левых и правых закручиваний, причем количество рядов, закрученных в одном направлении, и количество рядов, закрученных в другом направлении, суть два соседних числа Фибоначчи (см. рис. 7.7). (Если внимательно рассмотреть такую структуру от одного конца до другого, можно обнаружить «место перехода», где числа рядов сменяются на следующую пару соседних чисел Фибоначчи.) Любопытно, что гексагональный узор микротрубочек демонстрирует очень похожую особенность — в общем случае даже еще более точно, — причем состоит этот узор (по крайней мере, обычно) из 5 правых и 8 левых винтовых структур, как показано на рис. 7.8. На рис. 7.9 я попытался изобразить, как такие структуры могли бы «выглядеть» изнутри микротрубочки. Число 13 выступает здесь как общее количество витков в спирали: $5 + 8$. Любопытно также, что в двойных микротрубочках, встречающихся достаточно часто, внешний слой составной трубки обычно содержит 21 ряд димеров тубулина — следующее число Фибоначчи! (Не стоит, впрочем, чересчур увлекаться подобными построениями; например, в пучках микротрубочек в ресничках и центриолях бывает и по 9 рядов димеров — число, определенно *не принадлежащее* последовательности Фибоначчи.)

Откуда в структуре микротрубочек берутся числа Фибоначчи? Относительно еловых шишек, цветков подсолнечника и т. д. существует несколько вполне убедительных теорий — кстати, среди тех, кто серьезно занимался этим вопросом, был Алан Тьюринг (см. [198], с. 437). Однако к случаю микротрубочек эти теории, вполне возможно, неприменимы, и для такого уровня следует искать какие-то другие объяснения. Коруга [228] высказал предположение, что числа Фибоначчи в структуре микротрубочки повышают эффективность ее как «информационного процессора». В самом деле, согласно Хамероффу с коллегами (кото-

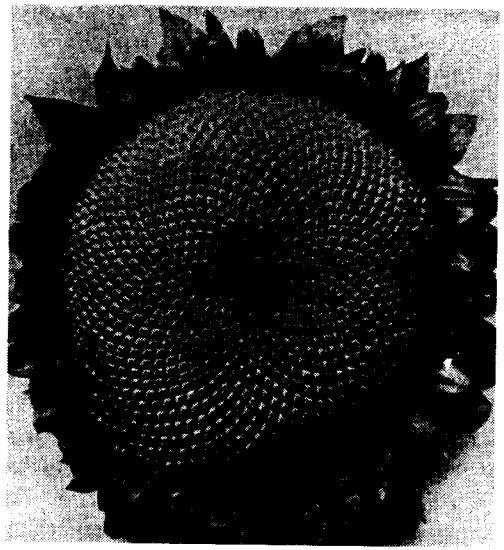


Рис. 7.7. Цветок подсолнечника. Как и во многих других растениях, отчетливо наблюдаются числа Фибоначчи. Во внешней области круга имеем 89 спиралей, закрученных по часовой стрелке, и 55 спиралей, закрученных против часовой стрелки. Ближе к центру появляются другие числа Фибоначчи.

рые пытаются нам это втолковать вот уже более десяти лет[4]), микротрубочки могут действовать как *клеточные автоматы*, передавая и обрабатывая сложные сигналы в виде волн различных состояний электрической поляризации молекул тубулина. Вспомним, что димеры тубулина могут существовать в двух (по крайней мере) различных конформационных состояниях и способных переходить из одного состояния в другое; последнее, очевидно, обусловливается сменой электрической поляризации молекулы на альтернативную. На состояние каждого димера воздействуют состояния поляризации каждого из шести его соседей (вследствие ван-дер-ваальсовых взаимодействий между ними), т. е. существуют вполне конкретные правила, определяющие конформацию каждого димера через конформации его соседей. Благодаря этому обстоятельству, каждая микротрубочка способ-



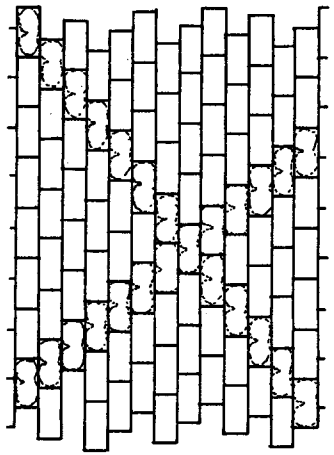


Рис. 7.8. Представим, что микротрубочка разрезана вдоль и затем развернута в полосу. Можно видеть, что молекулы тубулина располагаются вдоль наклонных линий, причем каждый новый виток смещен относительно предыдущего на 5 или 8 молекул (в зависимости от того, куда наклонена линия, вправо или влево).

на осуществлять передачу и обработку любого рода сообщений. С распространением сигналов, похоже, как-то связана транспортировка различных молекул вдоль микротрубочек, а также всевозможные соединения между соседними микротрубочками в виде своеобразных белковых «мостиков» — так называемые MAP (от *microtubule associated proteins*[6]); см. рис. 7.10. Коруга доказывает, что в случае структуры с числами Фибоначчи, подобной той, что реально наблюдается в микротрубочках, информация обрабатывается особенно эффективно. Должно быть, для такой организации микротрубочек и в самом имеется серьезная причина, поскольку, несмотря на некоторый разброс в числах, наблюдаемый в эукариотических клетках вообще, микротрубочки почти всех млекопитающих составлены именно из 13 рядов димеров.

Для чего микротрубочки нейронам? Каждый отдельный нейрон имеет свой цитоскелет. Какова его роль? Я уверен, что буду-

[6]Белки, ассоциированные с микротрубочками (англ.) — *Прим. перев.*

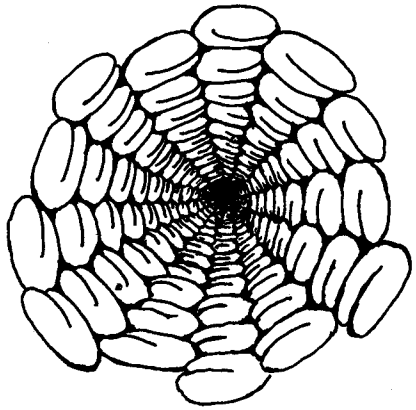


Рис. 7.9. Заглянем внутрь микротрубочки! Можно наблюдать спиральную структуру молекул тубулина 5 + 8.

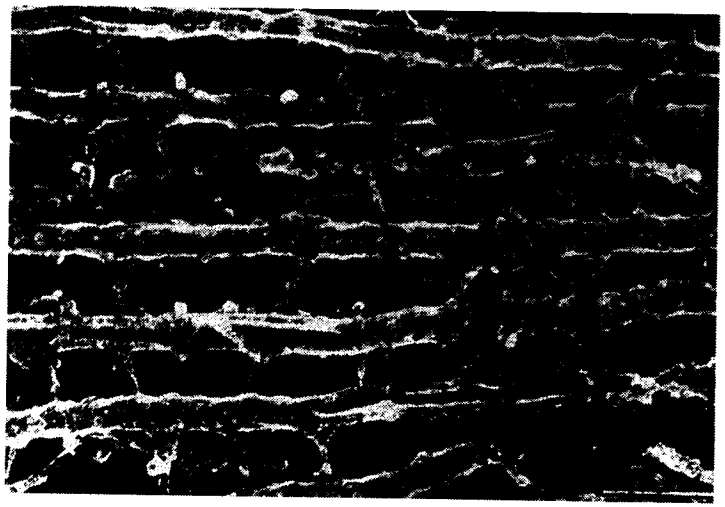


Рис. 7.10. Микротрубочки обычно соединяются друг с другом посредством «мостиков» из так называемых белков, ассоциированных с микротрубочками (MAP).



щим исследователям предстоит сделать в этой области еще немало открытий, однако кое-что мы знаем уже сейчас. В частности, микротрубочки нейронов могут быть очень и очень длинными (по сравнению с диаметром нейрона, который составляет лишь 25–30 нм) — до нескольких миллиметров или даже длиннее. Более того, в зависимости от обстоятельств они способны расти или сокращаться, а также транспортировать молекулы нейромедиаторов. Внутри аксонов и дендритов также имеются микротрубочки. Хотя, как правило, на всю длину аксона каждая отдельная микротрубочка не тянется, они образуют сообщающиеся сети, охватывающие всю клетку, соединяясь между собой посредством упоминавшихся выше MAP-мостиков. Микротрубочки, по-видимому, ответственны за поддержание интенсивности синапсов и, несомненно, за изменение этой интенсивности в случае необходимости. Более того, они, похоже управляют ростом новых нервных окончаний, направляя их к точкам соединений с другими нервными клетками.

Поскольку после окончательного формирования мозга деление нейронов прекращается, необходимости в этой функции центросомы здесь нет. В центросомах нейронов, расположенных вблизи ядра, часто вовсе нет центриолей. Микротрубочки тянутся от центросом к окрестности пресинаптических окончаний аксона, а также в другую сторону, к дендритам и, через сокращающиеся актиновые нити, к дендритным ши́пикам, часто образующим постсинаптические окончания синаптической щели 7.12. Эти ши́пики способны расти и вырождаться, что, по-видимому, является существенным элементом общей пластичности мозга, благодаря которой система взаимных соединений в мозге подвергается непрерывным тонким изменениям. Насколько мне известно, существуют убедительные экспериментальные свидетельства важной роли микротрубочек в управлении пластичностью мозга.

Упомянем еще об одном любопытном факте. В пресинаптических окончаниях аксонов содержатся некие ассоциированные с миктротрубочками вещества, «работа» которых связана с высвобождением нейромедиаторов, а молекулы весьма примечательны с геометрической точки зрения. Эти вещества — *клатрины* — строятся из белковых тримеров (так называемых клатриновых трискелионов), этаких полипептидных трехлучевых звезд. Объединяясь в молекулу клатрина, трискелионы образуют геометри-

чески правильные структуры, идентичные по общему строению многоатомным молекулам углерода, называемым «фуллеренами» (а также «бакиболами», или «мячами Баки»[7]) из-за их внешнего сходства со знаменитыми геодезическими куполами, которые проектировал и возводил американский архитектор Бакминстер Фуллер[5]. Клатрины, впрочем, гораздо больше фуллереновых молекул, поскольку одному атому углерода в фуллерене соответствует в клатрине целый трискелион, состоящий из нескольких аминокислот. Те клатрины, что заняты в высвобождении нейромедиаторов в синапсах, имеют форму *усеченного икосаэдра* — всем нам знакомого многогранника, по образу и подобию которого делают современные футбольные мячи (см. рис. 7.11 и 7.12).


Рис. 7.11. Молекула клатрина (похожая общей структурой на фуллерен, но составленная не из атомов углерода, а из более сложных субструктур — белковых тримеров, называемых трискелионами). Изображенный на рисунке клатрин напоминает внешне обыкновенный футбольный мяч.

В одном из предыдущих параграфов был поставлен важный вопрос: что управляет изменением интенсивности синапсов и определяет места размещения функционирующих синаптических связей? Учитывая имеющиеся свидетельства, можно уверенно предположить, что центральную роль в этих процессах играет *цитоскелет*. Как же это предположение может нам помочь в поиске невычислимой сущности разума? Пока что оно, похоже, говорит нам лишь о том, что потенциальная вычислительная мощность мозга оказывается гораздо большей, чем можно было бы

[7] Англ. *bucky balls*. — *Прим. перев.*



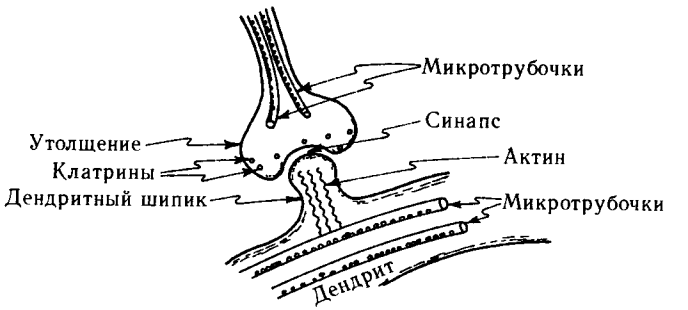



Рис. 7.12. Клатрины, подобные тому, что изображен на рис. 7.11, располагаются (вместе с окончаниями микротрубочек) в пресинаптическом утолщении аксона и, по всей видимости, участвуют в управлении интенсивностью синапса; также на интенсивность синапса влияют сокращающиеся актиновые нити в дендритных шипиках, управляемых микротрубочками.

ожидать, используй мозг в качестве простейших вычислительных блоков «цельные» нейроны.

В самом деле, если простейшими вычислительными блоками мы теперь будем считать димеры тубулина, то придется предположить, что потенциальная вычислительная мощность мозга просто неимоверно превосходит все то, что предполагали самые смелые теоретики от ИИ. Основываясь на «цельнонейронной» модели, Ханс Моравек в своей книге «Дети разума» [267] предположил, что человеческий мозг может в принципе достичь производительности порядка $10^{14}$ операций в секунду, но не более того; это при том, что в мозге имеется около $10^{11}$ функционирующих нейронов, каждый из которых способен посылать примерно по $10^3$ сигналов в секунду (см. § 1.2). Если же в качестве элементарного вычислительного блока взять димер тубулина, то следует учесть, что на каждый нейрон приходится около $10^7$ димеров; соответственно, элементарные операции теперь выполняются где-то в $10^6$ раз быстрее, в результате чего получаем $10^{27}$ операций в секунду. Возможно, производительность современных компьютеров и вправду уже начинает приближаться к первой цифре, $10^{14}$ операций в секунду (как весьма убе-

жденно доказывают Моравек и его единомышленники), однако несмотря на все эти успехи, достичь в обозримом будущем производительности $10^{27}$ операций в секунду не представляется возможным.

Разумеется, можно смело утверждать, что мозг работает далеко не со стопроцентной «микротрубочковой» эффективностью, какую приведенные выше цифры предполагают. Тем не менее, ясно, что возможность «микротрубочкового вычисления» (см. [183]) позволяет совсем по-иному взглянуть на некоторые из аргументов в пользу неминуемого наступления эпохи искусственного интеллекта человеческого уровня. Можем ли мы теперь поверить хотя бы в то, что уже сегодня возможно[6] численно воспроизвести умственную деятельность червя нематоды, только потому, что мы вроде бы «закартографировали» и численно смоделировали его нервную систему? Как было отмечено в § 1.15, умственные способности обычного муравья намного превосходят все то, что на настоящий момент реализовано посредством стандартных ИИ-процедур. Впору поинтересоваться, сколько же муравей выигрывает в производительности благодаря гигантскому массиву своих «микротрубочковых информационных нанопроцессоров», если сравнивать с тем, чего он смог бы добиться, располагай он лишь «переключателями цельнонейронного типа». Что до парамеции, то тут, как вы понимаете, оснований для предъявления иска нет.

Однако аргументы, представленные в первой части, предполагают гораздо более сильное заявление. Я утверждаю, что способность человека к пониманию выходит за рамки какой угодно вычислительной схемы. Если мозгом человека управляют микротрубочки, то в микротрубочковых процессах должно быть что-то принципиально отличное от простого вычисления. Я утверждал, что такая невычислимая активность должна быть следствием достаточно макроскопической квантовой когерентности, объединенной неким тонким образом с макроскопическим поведением — с тем, чтобы обеспечить возможность протекания в системе тех новых физических процессов, что придут на смену бытующей в современной физике паллиативной **R**-процедуре. В качестве первого шага мы должны выяснить, какова же подлинная роль *квантовой когерентности* в цитоскелетной активности.



## 7.5. Квантовая когерентность внутри микротрубочек

Есть ли у нас основания предполагать, что внутри микротрубочек существует квантовая когерентность? Вернемся ненадолго к обсуждавшимся в § 7.1 идеям Фрёлиха [131] о возможности феноменов квантовой когерентности в биологических системах. Он утверждал, что если энергия метаболической активности достаточно велика, а диэлектрические свойства задействованных в процессе материалов достаточно экстремальны, то существует возможность возникновения макроскопической квантовой когерентности, аналогичной той, что возникает в феноменах сверхпроводимости и сверхтекучести — иногда объединяемых общим термином *конденсация Бозе–Эйнштейна* — даже при относительно высоких температурах, какие, собственно, и характерны для биологических систем. Как выяснилось, не только метаболическая энергия достаточно велика, а диэлектрические свойства просто необыкновенно экстремальны (именно этот полученный в 1930-е годы поразительный экспериментальный результат и навел Фрёлиха на соответствующие размышления), но и имеется с некоторых пор даже прямое подтверждение предсказанных Фрёлихом внутриклеточных колебаний с частотой $10^{11}$ Гц [177].

В конденсате Бозе–Эйнштейна (который возникает еще и при работе лазера) большое количество частиц совместно образуют одно квантовое состояние. Это состояние описывается волновой функцией того же вида, что и в случае единичной частицы, — только здесь эта функция относится сразу ко всей совокупности образующих состояние частиц. Вспомним о непостижимой с классической точки зрения природе квантового состояния одной-единственной квантовой частицы (§§ 5.6, 5.11). В конденсате Бозе–Эйнштейна вся состоящая из множества частиц система ведет себя как одно целое, и ее квантовое состояние ничем не отличается от квантового состояния единичной частицы, меняется только масштаб. В этом увеличенном масштабе и возникает когерентность, при которой многие удивительные свойств квантовых волновых функций проявляются на макроскопическом уровне.

Первоначально Фрёлих полагал, что такие макроскопические квантовые состояния должны, скорее всего, возникать в

клеточных мембранах[8], однако теперь перед нами открывается другая (и, судя по всему, более правдоподобная) возможность: *микротрубочки*. Причем эта возможность, похоже, подтверждается экспериментально(7). Еще в 1974 году Хамерофф предположил [182], что микротрубочки могут действовать как «диэлектрические волноводы». Хочется верить, что Природа снабдила цитоскелетные структуры пустыми трубками отнюдь не просто так. Возможно, сами трубки обеспечивают эффективную изоляцию, позволяющую квантовому состоянию внутри трубки избегать сцепления с окружением в течение достаточно продолжительного времени. В этой связи интересно отметить, что Эмилио дель Джудиче и его коллеги из Миланского университета утверждали [79], что в результате квантового эффекта самофокусировки электромагнитных волн в цитоплазме клетки сигналы сосредотачиваются внутри области, диаметр которой не превышает внутреннего диаметра микротрубочки. Это может послужить еще одним подтверждением волноводной теории, однако возможно также, что этот эффект участвует в собственно образовании микротрубочек.

Тут имеется еще один интересный момент, и связан он с природой *воды*. Сами трубки, похоже, всегда остаются пустыми — факт сам по себе интересный и, возможно, значимый, особенно если учесть, что мы предполагаем найти внутри этих трубок управляемые условия, благоприятные для некоторого рода коллективных квантовых колебаний. «Пустые» в данном случае означает, что трубки по большей части заполнены просто водой (даже без растворенных в ней ионов). Можно было бы отметить, что «вода» (с характерным для жидкости беспорядочным движением молекул) вряд ли является образцом организованной структуры — во всяком случае достаточно организованной для возникновения в ней квантовокогерентных колебаний. Однако вода, содержащаяся в клетках, совсем не похожа на ту воду, которой заполнены океаны — неупорядоченное скопище несвязных, случайным образом движущихся молекул. Некоторая часть воды в

[8]Убежденным сторонником идеи, согласно которой конденсация Бозе — Эйнштейна способна привести к формированию того «отдельного самоощущения», которое можно счесть характерной особенностью сознания, является Иэн Маршалл [258], см. также [397], [398] и [243]. Ранее идею глобальных (существенно квантовых) макроскопических когерентных «голографических» процессов в мозге активно поддерживал Карл Прибрам [317, 318, 319].



клетках — какая именно часть, вопрос спорный — находится в *упорядоченном* состоянии (такую воду иногда называют «вицинальной», см. [183], с. 172). Такое упорядоченное состояние воды наблюдается на расстоянии до 3 нм от внешних поверхностей цитоскелета, иногда дальше. Представляется вполне разумным предположить, что вода остается упорядоченной и внутри микротрубочек, а это весьма благоприятствует возможности возникновения в этих трубках квантовокогерентных колебаний. (См., в частности, [213]).

Каким бы ни оказался окончательный статус этих захватывающих идей, одно мне совершенно ясно: вероятность того, что полностью классическое описание цитоскелета способно адекватно объяснить его поведение, ничтожно мала. С нейронами дело обстоит иначе, там описания в исключительно классическом духе и в самом деле представляются, по большому счету, вполне допустимыми. В самом деле, при ознакомлении с современными исследованиями цитоскелетных процессов бросается в глаза тот факт, что авторы то и дело прибегают к «помощи» квантовомеханических концепций, и я почти не сомневаюсь, что в будущем эта тенденция только усилится.

Впрочем, ясно также и другое: многие пока еще далеко не убеждены в том, что какие бы то ни было квантовые эффекты могут иметь столь непосредственное отношение к функционированию цитоскелета или мозга вообще. Даже если допустить, что работа микротрубочек и сознательная деятельность мозга суть прямой результат неких существенных эффектов квантовой природы, продемонстрировать эти самые эффекты посредством какого-нибудь убедительного эксперимента отнюдь не просто. Возможно, нам повезет, и удастся приспособить к микротрубочкам некоторые из стандартных процедур, которые применяются сегодня для демонстрации присутствия конденсатов Бозе — Эйнштейна в физических системах — например, при высокотемпературной сверхпроводимости. С другой стороны, может и не повезти — и тогда придется искать какие-то принципиально новые подходы. Возможно, нам удастся показать, что возбуждение микротрубочек предполагает ту же нелокальность, какую мы наблюдаем в ЭПР-феноменах (неравенства Белла и т. д., см. §§ 5.3, 5.4, 5.17), поскольку классического (локального) объяснения подобных эффектов не существует. Можно, например, выполнить измерения в двух точках одной микротрубочки (или

же разных микротрубочек) и получить результат, необъяснимый с точки зрения классической независимости событий в этих двух точках.

Каким бы ни было наше отношение к подобным предположениям, очевидно, что исследования микротрубочек еще даже не вышли из пеленок. И я нисколько не сомневаюсь, что они преподнесут нам в недалеком будущем множество потрясающих сюрпризов.

## 7.6. Микротрубочки и сознание

Есть ли прямые свидетельства того, что феномен *сознания* в той или иной мере обусловлен деятельностью цитоскелета и, в частности, его микротрубочек? Как ни странно, *есть*. Причем получено оно путем обращения к проблеме сознания с неожиданной стороны — с попытки выяснить, что может послужить причиной его *отсутствия*.

В поисках ответов на вопросы, касающиеся физических основ сознания, важную роль играет исследование причин и способов, весьма избирательно это самое сознание «отключающих». На такое способны, например, *препараты для общего наркоза*, причем это отключение абсолютно обратимо, главное — не превысить допустимую концентрацию. Замечательно то, что к общему наркозу приводит применение множества самых разных веществ, никак, казалось бы, не связанных друг с другом химически. К таким веществам относятся закись азота ($N_2O$), эфир ($CH_3CH_2OCH_2CH_3$), хлороформ ($CHCl_3$), галотан ($CF_3CHClBr$), изофлуран ($CHF_2OCHClCF_3$) и даже химически инертный (!) газ ксенон.

Если за общий наркоз «ответственна» не «химия», то что же тогда? Помимо химических взаимодействий, на молекулы действуют и другие силы, гораздо более слабые — например, так называемые *ван-дер-ваальсовы* силы. Силы Ван-дер-Ваальса — это слабое притяжение между молекулами, обладающими *электрическим дипольным моментом* («электрическим» эквивалентом магнитного дипольного момента, определяющего силу обычного магнита). Вспомним, что димеры тубулина могут находиться в двух различных конформациях. Конформации эти, по всей видимости, обусловлены тем, что в центре димера (в его «безводной» области) имеется электрон, который может занимать одно из двух возможных положений. От положения электрона зависит как общая форма диполя, так и его электрический момент. На способность молекул димера «переключаться» из одной конформации в другую влияют ван-дер-ваальсовы силы притяжения соседних молекул. Было высказано предположение [185], что действие анестезирующих веществ основано на ван-дер-ваальсовых взаимодействиях (в «гидрофобных» — водоотталкивающих — областях, см. [123]), которые препятствуют нормальным переключениям тубулина. Таким образом, как только анестезирующий газ просачивается в нервную клетку, его электрические дипольные свойства (которые вовсе не обязательно должны находиться в прямой зависимости от его химических свойств) останавливают работу микротрубочек. В общем и целом получается весьма правдоподобная картина действия общего наркоза. Ввиду очевидного отсутствия детального общепринятого описания действия анестетиков, достаточно логичной представляется точка зрения, согласно которой причиной потери сознания является ван-дер-ваальсово воздействие анестезирующих веществ на конформационную динамику белков мозга. Высока вероятность того, что такими белками являются именно димеры тубулина в микротрубочках нейронов — и что к потере сознания приводит именно обусловленное упомянутым воздействием прекращение функционирования микротрубочек.



В поддержку предположения, что общие анестетики воздействуют непосредственно на *цитоскелет*, отметим, что эти вещества «отключают» не только «высших животных», таких как млекопитающие и птицы. Точно так же (и примерно в тех же концентрациях) действует наркоз на парамеций, амеб и даже на зеленых слизевиков (что наблюдал Клод Бернар еще в 1875 году [27]). Подвергаются ли воздействию реснички парамеции или ее центриоль, в любом случае «поражается» какая-либо часть *цитоскелета*. Если мы допускаем, что поведением такого одноклеточного животного действительно управляет цитоскелет, то, во избежание противоречий, следует допустить и то, что анестезирующие вещества действуют именно на цитоскелет.

Я, разумеется, не утверждаю, что таких одноклеточных животных следует рассматривать как обладающих сознанием. Сознание — это совершенно иное дело. Вполне возможно, что для возникновения сознания, *помимо* должным образом функционирующих цитоскелетов, необходима еще куча самых разных ве-

щей. Я сейчас говорю лишь о том, что, согласно вышеприведенным рассуждениям, *без* работающего цитоскелета ни о каком сознании речь не может идти вообще. При прекращении функционирования системы цитоскелетов сознание мгновенно выключается — столь же мгновенно возвращаясь, как только функции цитоскелета восстанавливаются, при условии, что за прошедшее время не возникло каких-либо повреждений иного рода. Разумеется, нам по-прежнему не дает покоя вопрос, может ли в самом деле обладать некоей зачаточной формой сознания парамеция — или, коли уж на то пошло, отдельно взятая клетка человеческой печени — однако представленных соображений для ответа явно не достаточно. В любом случае, *форма* сознания должна самым фундаментальным образом определяться тонкой нейронной организацией мозга. Более того, если бы от этой организации ничего не зависело, то в нашей печени обитало бы ничуть не худшее сознание, чем в нашем мозге. Тем не менее, как недвусмысленно показывают представленные аргументы, важна *не только* нейронная организация мозга. Для наличия сознания жизненно необходима и цитоскелетная «начинка» этих самых нейронов.

Можно предположить, что для возникновения сознания в общем случае важен не сам цитоскелет *как таковой*, но некая *существенная физическая активность*, которую хитроумные биологи умудрились разглядеть в микротрубочковых процессах. Что же это за существенная физическая активность? Вся аргументация первой части книги подводила нас, в сущности, к простому выводу: если мы намерены подвести под процесс сознания физический фундамент, то нам понадобится нечто большее, чем численное моделирование. В предыдущих главах второй части мы успели договориться до того, что искать это большее следует на границе между квантовым и классическим уровнями, как раз там, где современная физики предлагает (за неимением лучшего) воспользоваться процедурой **R**, а я настаиваю на разработке *новой* физической теории — теории процедуры **OR**. В настоящей главе мы попытались отыскать в мозге такое место, где квантовые процессы могли бы определять классическое поведение, и, похоже, пришли к выводу, что этот квантово-классический интерфейс осуществляет фундаментальное воздействие на поведение мозга посредством *цитоскелетного управления интенсивностью синаптических связей*. Попробуем рассмотреть эту картину более основательно.



## 7.7. Модель разума

Как уже отмечалось в § 7.1, мы вполне можем согласиться с тем, что сами по себе нервные сигналы можно рассматривать как исключительно классические феномены, — особенно если предположить, что такие сигналы настолько возмущают окружение, что квантовая когерентность на этом этапе не может сохраняться сколько-нибудь долго. Допустим далее, что синаптические связи и их интенсивность всегда остаются неизменными; в этом случае воздействие любого возбужденного нейрона на следующий нейрон также поддается классическому описанию — за исключением, впрочем, случайной составляющей, которая появляется на этом этапе. Активность мозга в таких условиях целиком и полностью вычислима, т. е. *в принципе* возможно построить его численную модель. Это не значит, что такая модель будет в точности имитировать деятельность того *конкретного* мозга, схема синаптических связей которого совпадает со схемой модели (вследствие наличия упомянутых случайных составляющих), однако модель сможет воспроизвести *типичную* активность такого мозга и, как следствие, предсказать типичное поведение того или иного индивидуума, этим мозгом управляемого (см. § 1.7). Более того, утверждение это носит по большей части чисто *принципиальный* характер. Ничто не указывает на то, что при современном уровне развития технологий такую численную модель действительно можно построить. Я также предполагаю, что случайные составляющие *подлинно* случайны. Возможность привлечения дуалистического внешнего «разума» с целью воздействия на упомянутые случайности здесь не рассматривается вовсе (см. § 1.7).

Таким образом, получаем (по крайней мере, предварительно), что при условии *постоянства* синаптических связей мозг действительно работает как своего рода *компьютер* — пусть и со встроенными случайными составляющими. Как мы показали в первой части, в высшей степени невероятно, чтобы такая схема могла когда-либо послужить основой для построения модели человеческого сознательного понимания. С другой стороны, если специфические синаптические связи, определяющие данный конкретный нейронный компьютер, постоянно меняются, а управление этими изменениями возложено на некий *невычислимый*

процесс, то вполне возможно, что такая расширенная модель действительно окажется способна воспроизвести поведение осознающего себя мозга.

Что же это может быть за невычислимый процесс? Здесь следует вспомнить о *глобальной* природе сознания. Если, скажем, взять $10^{11}$ независимых цитоскелетов, каждый из которых внесет в общее дело свою невычислимую долю, то пользы от этого нам будет немного. Согласно аргументам первой части, невычислимое поведение и в самом деле неразрывно связано с процессом сознания — по крайней мере, настолько, чтобы можно было определенно утверждать, что *некоторые* проявления сознания, прежде всего способность *понимать*, невычислимы в принципе. Однако это не имеет никакого отношения ни к отдельным цитоскелетам, ни к отдельным микротрубочкам внутри цитоскелета. Никто в здравом уме не станет предполагать, что вот этот цитоскелет или вот та микротрубочка в состоянии хоть что-нибудь «понять» в рассуждениях Гёделя! Понимание работает в гораздо более глобальном масштабе, и если в процессе каким-то образом участвуют цитоскелеты, то этот феномен должен носить коллективный характер, задействуя огромное количество цитоскелетов одновременно.

Согласно Фрёлиху, биологические макроскопические коллективные квантовые феномены — может быть, той же природы, что и конденсат Бозе–Эйнштейна, — определенно возможны, даже внутри «горячего» мозга (см. также [258]). Здесь же мы предполагаем, что в относительно «крупных» квантовокогерентных состояниях должны участвовать не только молекулы внутри отдельных микротрубочек — такое состояние должно распространяться от одной микротрубочки к другой. Квантовая когерентность должна не просто «охватить» одну-единственную микротрубочку (пусть и, как мы помним, весьма протяженную), но перейти дальше, в результате чего большое количество различных микротрубочек в цитоскелете нейрона — если не все — должны образовать единое квантовокогерентное состояние. Мало того, квантовая когерентность должна преодолеть «синаптический барьер» между «своим» нейроном и следующим. Не много проку в глобальности, которая разбросана по изолированным друг от друга клеткам! Самостоятельная единица сознания может возникнуть, в нашем описании, лишь тогда, когда квантовая когерентность в том или ином виде получает возможность распространяться на некую существенную (по меньшей мере) часть всего мозга.



И вот такое вот поразительное — я бы даже сказал, почти невероятное — устройство Природе пришлось создавать с помощью одних лишь биологических средств. Я, впрочем, убежден (и не без оснований), что у нее таки все получилось, и главным свидетельством тому может служить факт наличия у нас разума. Нам еще многое предстоит понять в биологических системах и в том, как они творят свои чудеса — многое в биологии далеко превосходит возможности современных физических технологий. (Взять, к примеру, крохотного, в миллиметр величиной, паучка, искусно плетущего замысловатую паутину.) Вспомним и об экспериментах Аспекта (см. §5.4), в которых наблюдались (с помощью вполне *физических* устройств) кое-какие квантовокогерентные эффекты (ЭПР-сцепленность пар фотонов), действующие на расстоянии нескольких метров. Несмотря на технические трудности, связанные с проведением экспериментов, позволяющих обнаружить такие «дальнодействующие» квантовые эффекты, не следует исключать возможность, что Природа смогла отыскать биологические способы как для этого, так и для чего-нибудь еще. Присущую жизни «изобретательность» нельзя недооценивать.

Как бы то ни было, представляемые мною аргументы предполагают не только макроскопическую квантовую когерентность. Они предполагают, что биологическая система, называемая человеческим мозгом, каким-то образом ухитрилась воспользоваться в своих интересах физическими феноменами, человеческой же физике неизвестными! Эти феномены когда-нибудь опишет несуществующая пока теория **OR**, которая свяжет вместе классический и квантовый уровни и, я убежден, заменит временную **R**-процедуру иной, чрезвычайно тонкой и невычислимой (но все же, несомненно, математической) физической схемой.

То, что физики-люди, по большей части, пока еще ничего не знают о вышеупомянутой несуществующей теории, разумеется, не может заставить Природу отказаться от ее применения в своих биологических построениях. Она пользовалась принципами ньютоновской динамики задолго до Ньютона, электромагнитными феноменами задолго до Максвелла и квантовой механикой задолго до Планка, Эйнштейна, Бора, Гейзенберга, Шрёдингера и Дирака — в течение нескольких миллиардов лет! Лишь по

причине свойственной нашему веку нелепой самонадеянности столь многие сегодня пребывают в уверенности, что нам известны все фундаментальные принципы, лежащие в основе каких угодно тонких биологических процессов. Когда какой-нибудь живой организм по счастливой случайности натыкается на такой тонкий процесс, он начинает его активно применять и, возможно, получает в результате некие преимущества перед своими менее удачливыми соседями. Тогда Природа благословляет этот организм вместе со всеми его потомками и позволяет новому тонкому физическому процессу сохраниться в последующих поколениях — посредством, например, такого мощного инструмента, как естественный отбор.

Когда появились первые эукариотические клетки-животные, они, должно быть, обнаружили, что наличие у них примитивных микротрубочек дает им огромные преимущества. В результате возникло (посредством тех самых процессов, о которых мы здесь говорим) некое организующее воздействие, которое, возможно, привело к развитию зачатков способности к своего рода целенаправленному поведению, что помогло им выжить и вытеснить лишенных микротрубочек конкурентов. Называть такое воздействие «разумом», конечно же, еще рано; и все же оно возникло, как я полагаю, благодаря некоему тонкому пограничному взаимодействию между квантовыми и классическими процессами. Тонкостью же своей это взаимодействие обязано хитроумному физическому процессу **OR** — по-прежнему в подробностях нам неизвестному, — который в условиях не столь тонкой организации принимает вид того грубого квантовомеханического **R**-процесса, которым мы пока за неимением лучшего пользуемся. Далекие потомки тех клеток-животных — нынешние парамеции и амебы, а также муравьи, лягушки, цветы, деревья и люди — сохранили преимущества, которыми этот хитроумный процесс одарил древних эукариотов, и добавили новые, отвечающие новым многочисленным и самым разнообразным целям. Только будучи наложен на высокоразвитую нервную систему, этот процесс оказался, наконец, в состоянии реализовать свой гигантский потенциал — дав начало тому, что мы, теперь уже с полным правом, называем «разумом».

Итак, мы допускаем, что в глобальной квантовой когерентности может участвовать вся совокупность микротрубочек в цитоскелетах большого семейства нейронов мозга — или, по



крайней мере, что между состояниями различных микротрубочек в мозге наличествует достаточная квантовая сцепленность, — т. е. полностью *классическое* описание коллективного поведения этих микротрубочек *невозможно*. Можно представить, что в микротрубочках возникают сложные «квантовые колебания» — там, где изоляции, обеспечиваемой самими трубками, достаточно для того, чтобы квантовая когерентность сохранялась хотя бы частично. Велик соблазн предположить, что «клеточноавтоматные» вычисления, которые, по мнению Хамероффа и его коллег, должны выполняться *на поверхности* трубок, могут оказаться связанными с предполагаемыми квантовыми колебаниями *внутри* трубок (например, теми, что описаны в [79] или в [213]).

Заметим в этой связи, что частота, предсказанная Фрёлихом для коллективных квантовых колебаний (и подтвержденная наблюдениями Грундлера и Кайльмана [177]) — порядка $5 \times 10^{10}$ Гц (т. е. $5 \times 10^{10}$ колебаний в секунду), — практически совпадает с частотой, с которой, по Хамероффу, димеры тубулина в микротрубочковых клеточных автоматах «переключаются» из одного состояния в другое. Таким образом, если внутри микротрубочек и в самом деле работает фрёлихов механизм, то следует признать, что какая-то связь между этими двумя типами активности действительно имеется[9].

Впрочем, если бы такая связь была слишком сильной, то квантовый характер внутренних колебаний неизбежно означал бы, что и вычисления на поверхности самих трубок необходимо рассматривать квантовомеханически. Иначе говоря, на поверхности микротрубочек происходили бы самые настоящие *квантовые вычисления* (см. § 7.3)! Следует ли воспринимать такую возможность всерьез?

Трудность заключается в том, что для таких вычислений, по-видимому, необходимо, чтобы изменения конформаций димеров не возмущали сколько-нибудь заметным образом молеку-

[9] Гораздо менее понятно, впрочем, существует ли сколько-нибудь прямая связь между упомянутыми сравнительно высокочастотными процессами и более привычной «волновой» активностью мозга (например, альфа-ритмом с частотой 8–12 Гц). Предполагается лишь, что такие низкие частоты могут возникать как «частоты биений», однако никакой связи пока не установлено. Особо примечательными в этой связи представляются не так давно обнаруженные колебания с частотой 35–75 Гц, ассоциирующиеся, по-видимому, с областями мозга, ответственными за сознательное внимание. Колебания эти, похоже, обладают какими-то загадочными нелокальными свойствами. (См. [107], [167], [64], [65] и [63]).

лы окружения. Здесь уместно вспомнить о том, что окружающая микротрубочку область заполнена водой в *упорядоченном* состоянии, прочие же вещества в эту область не допускаются (см. [183], с. 172), что в совокупности может обеспечить некоторое квантовое экранирование. С другой стороны, микротрубочки соединены друг с другом «мостиками» MAP (см. § 7.4) — причем по некоторым из них производится транспорт разных «посторонних» молекул, — и передача сигналов вдоль трубок (см. [183], с. 122) не может на эти мостики не воздействовать. Из этого последнего факта вполне недвусмысленно следует, что «вычисления», которыми занята трубка, могут и в самом деле возмутить окружение до такой степени, что их поневоле придется рассматривать классически. Интенсивность возмущения невелика ввиду малости перемещаемых масс (по **OR**-критерию, предложенному в § 6.12), однако для того, чтобы вся система продолжала оставаться на квантовом уровне, необходимо, чтобы эти возмущения не проникали внутрь клетки и не распространялись далее, за ее пределы. На мой взгляд, неопределенности здесь (как в отношении реальной физической ситуации, так и в отношении применимости к ней критерия **OR** из § 6.12) остается вполне достаточно для того, чтобы помешать нам решить, уместен на данном этапе чисто классический подход или нет.

Как бы то ни было, предположим, в рамках настоящего рассуждения, что микротрубочковые вычисления следует рассматривать как существенно классические — в том смысле, что мы не ожидаем, что квантовые суперпозиции различных вычислений играют здесь сколько-нибудь значимую роль. С другой стороны, допустим, что *внутри* трубок имеют место подлинно квантовые колебания некоего рода, причем между внутренними квантовыми и внешними классическими свойствами каждой трубки существует некая тонкая связь. Согласно такой картине, именно в этом тонком взаимодействии существенно проявляются неизвестные пока правила искомой новой теории **OR**. Внутренние квантовые «колебания» должны определенным образом воздействовать на внешние вычисления на трубках, однако в этом нет ничего нелогичного — учитывая те механизмы, которые, как мы предполагаем, ответственны за клеточноавтоматное поведение микротрубочек (слабые взаимодействия ван-дер-ваальсова типа между соседними димерами тубулина).



В результате мы получаем картину некоего глобального квантового состояния, которое когерентно объединяет процессы внутри трубок и в котором участвует вся совокупность микротрубочек в той или иной обширной области мозга. Это состояние (которое вовсе не обязательно является просто «квантовым состоянием» в том традиционном смысле, который вкладывает в это понятие стандартный квантовый формализм) также некоторым образом воздействует на вычисления, выполняемые на микротрубочках, — для точного описания такого воздействия понадобится гипотетическая невычислимая **OR**-физика, которой у нас пока нет, но которая, я убежден, нам крайне необходима. «Вычислительная» активность конформационных изменений молекул тубулина управляет транспортом молекул вдоль наружной поверхности микротрубочек (см. рис. 7.13) и в конечном итоге воздействует на интенсивность синапса в его пре- и постсинаптических окончаниях. Таким образом, через посредство *внешних* вычислений, когерентная квантовая организация *внутри* микротрубочек способна влиять на изменения в синаптических связях нейронного компьютера в текущий момент.

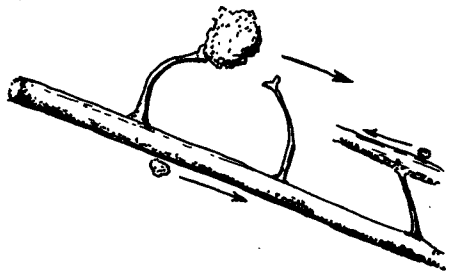


Рис. 7.13. Мостики MAP, помимо прочего, транспортируют крупные молекулы, тогда как меньшие молекулы перемещаются непосредственно вдоль микротрубочек.

Такая картина открывает простор для самых различных умозрительных построений. Например, можно отвести в ней некую роль нелокальности ЭПР-эффектов квантовой сцепленности. Определенную роль может играть и квантовая контрфактуальность. Представим, что нейронный компьютер готов выполнить некое вычисление, которое он в действительности не выполняет,

но (как в случае задачи об испытании бомб) сам факт того, что он *может* это вычисление выполнить, вызывает эффект, отличный от того, который имел бы место, не будь у компьютера такой возможности. Таким образом, классическая «схема соединений» нейронного компьютера в любой момент времени может воздействовать на внутреннее цитоскелетное состояние, даже если возбуждение нейронов, активирующее данную конкретную «схему», в действительности не происходит. Можно еще поразмышлять над возможными аналогами такого рода феноменов в каких-либо более привычных умственных занятиях, каким мы то и дело предаемся, но мне почему-то кажется, что углубляться в обсуждение этих занятий здесь не стоит.

Согласно предлагаемой мною предварительной точке зрения, сознание есть проявление такого квантовосцепленного внутреннего состояния цитоскелета вкупе с участием этого состояния во взаимодействии (**OR**) между процессами квантового и классического уровней. Компьютерообразная система нейронов, классическим образом соединенных друг с другом, непрерывно подвергается воздействию упомянутых цитоскелетных процессов, выступающих в роли проявлений «свободы воли» (что бы мы под этими словами ни понимали). Нейроны в этой системе выполняют функции, скорее, *увеличительных стекол*, посредством которых микроскопические цитоскелетные процессы «поднимаются» на уровень, на котором возможно воздействие на другие органы тела — например, на мышцы. Соответственно, нейронный уровень описания, к которому сводится модное нынче представление о мозге и разуме, является не более чем *тенью* цитоскелетных процессов более глубокого уровня — именно там, в глубине, находится физический фундамент *разума*, который мы столь упорно разыскиваем!

Эта картина, надо признать, не лишена некоторой умозрительности, однако она ни в чем не противоречит современным научным представлениям. В предыдущей главе мы убедились, что есть весьма веские причины (основанные на соображениях, не выходящих за рамки сегодняшней физики) полагать, что эта самая физика нуждается в серьезном пересмотре — для того, чтобы объяснять и описывать новые эффекты на том же уровне, на котором, по-видимому, происходят процессы в микротрубочках и, возможно, на границе цитоскелет/нейрон. Согласно представленным в первой части аргументам, для отыскания физического



«обиталища» сознания необходимо «расчистить» в физике место для невычислимых физических процессов, единственная же приемлемая возможность такой расчистки заключается, как я показываю уже во второй части, в последовательном замещении редукции квантового состояния, обозначенной здесь буквой **R**, новой, объективной редукцией **OR**. Теперь мы должны ответить на вопрос, есть ли какие-нибудь чисто *физические* основания ожидать, что процедура **OR** действительно окажется в принципе невычислимой. Как вскоре выяснится, некоторые основания такого рода, учитывая сделанные в § 6.12 предположения, действительно имеются.

## 7.8. Невычислимость в квантовой гравитации (1)

Ключевым требованием предшествующих рассуждений было то, что какой бы новый физический процесс ни пришел на смену вероятностной **R**-процедуре, применяемой в стандартной квантовой теории, его неотъемлемым свойством должна быть того или иного рода невычислимость. В § 6.10 я показал, что этот новый физический процесс, **OR**, должен сочетать в себе принципы квантовой теории с принципами общей теории относительности Эйнштейна — т. е. представлять собой квантово-*гравитационный* феномен. Есть ли какие-нибудь свидетельства в пользу того, что невычислимость может оказаться существенным свойством той теории (какой бы она ни была), которая в конечном счете корректно объединит (надлежащим образом модифицировав) квантовую теорию и общую теорию относительности?

Исследуя квантовую гравитацию, Роберт Герох и Джеймс Хартл столкнулись однажды с численно неразрешимой проблемой — *проблемой топологической эквивалентности четырехмерных многообразий* [144]. В основном их занимал вопрос о том, как определить, что два данных четырехмерных пространства «одинаковы» с топологической точки зрения (т. е. одно из этих пространств посредством непрерывной деформации можно довести до полного совпадения с другим пространством, причем деформация эта не допускает каких бы то ни было разрывов или слияний пространств). На рис. 7.14 топологическая эквивалентность проиллюстрирована на примере двухмерного случая, где

мы видим, что поверхность чашки топологически одинакова с поверхностью кольца, но отлична от поверхности шара. В двухмерном случае проблема топологической эквивалентности разрешима вычислительным путем, в случая же *четырех* измерений, как показал в 1958 году А. А. Марков [256], алгоритма для решения такой задачи не существует. Более того, доказательство Маркова эффективно демонстрирует, что если бы такой алгоритм существовал, то его можно было бы преобразовать в алгоритм, позволяющий решить *проблему остановки*, т. е. найти способ определять, завершится в той или иной ситуации работа машины Тьюринга или нет. Поскольку, как мы выяснили в § 2.5, такого алгоритма не существует, значит, не может быть и алгоритма для решения проблемы эквивалентности четырехмерных многообразий.

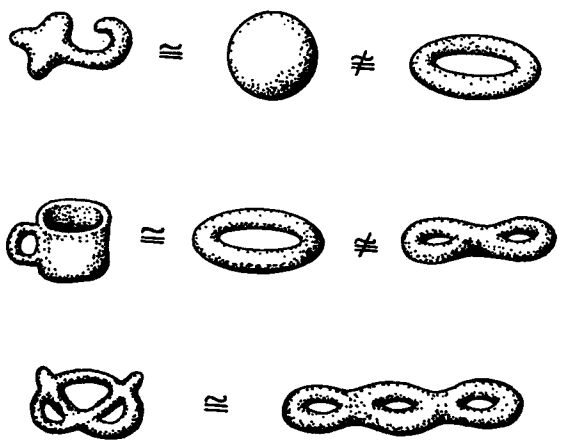


Рис. 7.14. Двухмерные замкнутые поверхности, которые можно классифицировать численно (грубо говоря, путем подсчета количества «ручек»). Четырехмерные же замкнутые «поверхности» численно классифицировать *невозможно*.

Существует множество других классов математических задач, которые неразрешимы численно. Две из них — десятую проблему Гильберта и задачу о замощении — мы обсуждали в § 1.9.



Еще один пример — задачу со словами (для полугрупп) — можно найти в НРК, с. 130–132.

Следует пояснить, что термин «численно неразрешимый» не означает, что в данном классе имеются отдельные задачи, которые невозможно решить в принципе. Он означает лишь то, что не существует систематического (алгоритмического) способа решить *все* задачи этого класса. В том или ином отдельном случае порой оказывается возможным получить решение благодаря человеческой находчивости и проницательности, подкрепленной, может быть, некоторыми вычислениями. *Может*, напротив, случиться и так, что решение каких-то задач из класса окажется человеку не по силам (даже если он возьмет в помощники машину). Похоже, никто об этом феномене ничего определенного не знает, поэтому каждый волен составлять обо всем этом свое собственное мнение. Впрочем, как вполне *недвусмысленно* показывает «гёделевско-тьюринговское» рассуждение из § 2.5 (вкупе с аргументацией главы 3), задачи таких классов, *доступные* человеческому пониманию и проницательности (подкрепленным вычислениями, если хотите), все равно образуют класс, который численно неразрешим. (Для проблемы остановки, например, в § 2.5 показано, что класс вычислений, незавершаемость которых в состоянии установить человек, невозможно охватить каким-либо познаваемо обоснованным алгоритмом $A$ — а от этого уже отталкиваются аргументы главы 3.)

Что касается Героха, Хартла и квантовой гравитации, то проблема эквивалентности четырехмерных многообразий проникла в их анализ постольку, поскольку, согласно *стандартным* правилам квантовой теории, квантово-гравитационное состояние предполагает суперпозиции (с комплексными весовыми коэффициентами) всех возможных геометрий — *пространственно-временны́х*, в данном случае, геометрий, т. е. четырехмерных объектов. Для того чтобы понять, как определять такие суперпозиции каким-либо уникальным образом (во избежание путаницы при подсчете), необходимо знать, какие пространства-времена считать различными, а какие — одинаковыми. Проблема топологической эквивалентности представляет собой, таким образом, лишь часть более обширной задачи.

Читатель спросит: если вдруг подход Героха – Хартла к квантовой гравитации окажется физически корректным, будет ли это означать, что эволюция физических систем включает в себя нечто

существенно невычислимое? Вряд ли на этот вопрос можно дать ясный и однозначный ответ. Мне не ясно даже, так ли непременно из численной неразрешимости проблемы топологической эквивалентности следует неразрешимость более полной проблемы геометрической эквивалентности. Мне не ясно также, какое отношение этот подход может иметь (если вообще может) к искомой объективной редукции, которая предполагает изменения в самой структуре собственно квантовой теории, связанные с необходимостью учета гравитационных эффектов. Тем не менее, работа Героха — Хартла и в самом деле вполне определенно указывает на то, что невычислимость может-таки сыграть свою роль в окончательной, физически корректной теории квантовой гравитации.

## 7.9. Машины с оракулом и физические законы

Можно, впрочем, задать и иной вопрос. Предположим, что новая теория квантовой гравитации действительно окажется невычислимой теорией — в том, в частности, смысле, что она позволит нам сконструировать физическое устройство, способное решить проблему остановки. Будет ли этого достаточно для разрешения всех проблем, порожденных нашими размышлениями о доказательстве Гёделя — Тьюринга в первой части книги? Как ни удивительно, ответ — *нет*!

Попробуем разобраться, почему способность решить проблему остановки ничем нам не поможет. В 1939 году Тьюринг предложил одну важную концепцию, имеющую к этому вопросу самое непосредственное отношение, — концепцию *оракула*. Идея такова: оракул есть нечто (предположительно, воображаемый объект, существующий лишь в голове самого Тьюринга и вовсе не обязательно реализуемый физически), что действительно может решить проблему остановки. Так, если дать оракулу пару натуральных чисел $q$ и $n$, то он через некоторое конечное время выдаст нам ответ **ДА** или **НЕТ**, в зависимости от того, завершится в конце концов вычисление $C_q(n)$ или нет (см. §2.5). В §2.5 мы доказываем вывод Тьюринга о том, что такой оракул, действующий исключительно вычислительными методами, создать невозможно, однако там ничего не говорится о том, что оракул невозможно построить физически. Чтобы прийти к такому выводу, мы должны твердо знать, что физические законы являются



по своей природе вычислительными — а мы этого не знаем, о чем, собственно, и идет, главным образом, речь во второй части. Следует также отметить, что физическая возможность создания оракула не является, насколько я могу судить, следствием из той точки зрения, которую я здесь отстаиваю. Как уже упоминалось, никто не требует, чтобы все проблемы остановки были доступны человеческому пониманию и проницательности, поэтому нет никаких оснований и полагать, что некое физически реализуемое устройство непременно справится со всеми этими проблемами своей физической реализуемости.

В дальнейшем обсуждении Тьюринг рассмотрел модификацию понятия вычислимости, когда оракула можно вызвать на любом желаемом этапе вычисления. Таким образом, *машина с оракулом* (выполняющим *оракул-алгоритм*) представляет собой самую обыкновенную машину Тьюринга, только к ее стандартным вычислительным операциям добавлена еще одна: «Вызвать оракул и спросить у него, завершается ли вычисление $C_q(n)$; по получении ответа продолжать вычисление, учитывая полученный ответ». Оракул можно вызывать снова и снова, если появляется такая необходимость. Отметим, что машина с оракулом является точно таким же детерминированным объектом, как и обычная машина Тьюринга (это для иллюстрации того факта, что вычислимость и детерминизм суть совершенно разные вещи). В принципе, вселенная, которая функционирует детерминированно как машина с оракулом, точно так же возможна, как и вселенная, которая функционирует детерминированно как машина Тьюринга. («Игрушечные вселенные», описанные в § 1.9 и в НРК, на с. 170, представляют собой, по сути, вселенные-машины-с-оракулом.)

Может ли оказаться так, что и наша собственная Вселенная функционирует как машина с оракулом? Любопытно, что с помощью приведенных в первой части книги аргументов оракул-машинная модель математического понимания «развенчивается» столь же успешно, как и аналогичная модель на основе машины Тьюринга, причем изменений почти не требуется. Нужно всего лишь взять доказательство из §2.5 и условиться, что запись «$C_q(n)$» обозначает теперь «выполнение $q$-й машиной с оракулом действия над натуральным числом $n$». Впрочем, лучше ввести другое обозначение, скажем, $C'_q(n)$. Как и в случае обычных машин Тьюринга, мы можем составить (вычислимым образом) пронумерованный список машин с оракулом. Что касается

их спецификаций, единственной дополнительной особенностью является то, что мы должны, помимо прочего, учитывать, на каких этапах вычисления вызывается оракул; никакой новой проблемы такой учет не составит. Далее мы заменяем *алгоритм* $A(q, n)$ из §2.5 *оракул-алгоритмом* $A'(q, n)$, который, в соответствии с исходным допущением, олицетворяет собой всю совокупность доступных человеческому пониманию и человеческой проницательности средств, необходимых для однозначного установления факта незавершаемости операции $C'_q(n)$ оракула. В точности повторяя доказательство, приходим к следующему выводу:

> $\mathscr{G}'$ Для установления математической истины математики не применяют заведомо обоснованные оракул-алгоритмы.

Отсюда следует неутешительное заключение: физический процесс, функционирующий как машина с оракулом, наших проблем также не решит.

Вообще говоря, весь процесс можно повторить, применив его к «машинам с оракулом второго порядка», которым позволяется вызывать при необходимости *оракул второго порядка* — который способен установить, завершится работа обычной машины с оракулом или нет. Как и в предыдущем случае, приходим к выводу:

> $\mathscr{G}''$ Для установления математической истины математики не применяют заведомо обоснованные оракул-алгоритмы второго порядка.

Очевидно, что этот процесс можно повторять снова и снова — подобно многократной гёделизации, описанной нами в связи с возражением **Q19**. Для каждого рекурсивного (вычислимого) ординала $\alpha$ вводится концепция машины с оракулом $\alpha$-го порядка, и мы снова получаем все тот же вывод:

> $\mathscr{G}^\alpha$ Для установления математической истины математики не применяют заведомо обоснованные оракул-алгоритмы $\alpha$-го порядка, где $\alpha$ — любой вычислимый ординал.

Окончательное следствие из всего этого несколько даже пугает. Получается, что нам предстоит отыскать невычислимую физическую теорию, способную заглянуть дальше, чем описание машин с оракулом любого вычислимого уровня (или, возможно, еще дальше).



Нисколько не сомневаюсь, что найдутся читатели, которые скажут, что вот уж тут-то мои рассуждения окончательно растеряли последние крохи правдоподобия, которые в них еще оставались! И, разумеется, такие чувства вполне понятны. Непонятно лишь нежелание хотя бы ознакомиться со всеми доказательствами, которые я уже в подробностях приводил ранее. Нужно просто вновь пройти по всем доказательствам в главах 2 и 3, заменяя в них машины Тьюринга на машины с оракулом $\alpha$-го порядка. Не думаю, что такая замена как-то существенно повлияет на суть этих доказательств, но меня, если честно, приводит в содрогание перспектива только ради нее повторять их здесь заново. Следует, впрочем, указать на еще одно обстоятельство: нет никакой необходимости в том, чтобы человеческое понимание приобрело ту же мощь, что и *какая угодно* машина с оракулом. Как было отмечено выше, вывод $\mathscr{G}$ *вовсе* не обязательно предполагает, что человеческого понимания, в принципе, достаточно для того, чтобы решить любой конкретный случай проблемы остановки. Таким образом, все это не означает, что искомые физические законы в принципе должны непременно оказаться, более общими, нежели те, которыми описываются машины с оракулом любого вычислимого уровня (или хотя бы первого). Нам нужно лишь отыскать нечто, не являющееся эквивалентом *любой* конкретной машины с оракулом (включая сюда и машины с оракулом *нулевого* уровня, т. е. собственно машины Тьюринга). Возможно, эти физические законы опишут нечто просто-напросто *иное*.

## 7.10. Невычислимость в квантовой гравитации (2)

Вернемся к квантовой гравитации. Необходимо подчеркнуть, что в настоящее время общепринятой теории квантовой гравитации не существует — нет даже сколько-нибудь приемлемых кандидатов. Есть зато множество самых разных и порой совершенно восхитительных гипотез[8]. Та, которую я хочу сейчас представить, требует, как и подход Героха–Хартла, учета квантовых суперпозиций различных *пространств-времен*.

(Многие гипотезы говорят лишь о суперпозициях трехмерных пространственных геометрий, что несколько отличается.) Предположение (за авторством Дэвида Дойча[9]) заключается в том, что в суперпозициях должны участвовать не только «правильные» пространственно-временны́е геометрии, в которых время ведет себя достаточно благоразумно, но и «неправильные» пространства-времена, в которых имеются *замкнутые времениподобные линии*. Такое пространство-время представлено на рис. 7.15. *Времениподобная линия* описывает возможную историю частицы (классической), а «времениподобной» она называется потому, что во всех точках локального светового конуса линия всегда направлена внутрь конуса, т. е. локальная абсолютная скорость не превышается — в соответствии с требованием теории относительности (см. §4.4). Смысл *замкнутости* времениподобной линии в том, что мы можем представить себе «наблюдателя»[10], для которого такая линия является мировой линией, т. е. линией, описывающей в данном пространстве-времени историю его собственного тела. Такой наблюдатель по прошествии некоторого конечного времени (согласно его восприятию) окажется в своем прошлом (перемещение во времени!). У него появляется возможность сделать что-нибудь такое (при условии, что он обладает какой-никакой «свободой воли»), чего он раньше никогда не делал, что неизбежно ведет к противоречию. (Обычно в таких умопостроениях наблюдатель убивает собственного дедушку «прежде», чем на свет появится его же отец — или совершает что-нибудь еще столь же волнительное.)

Рассуждения такого рода сами по себе являются достаточной причиной для того, чтобы не воспринимать пространства-времена с замкнутыми времениподобными линиями всерьез — в качестве возможных моделей реально существующей классической Вселенной. (Любопытно, что первым модель пространства-времени с замкнутыми времениподобными линиями предложил в 1949 году не кто иной, как Курт Гёдель. Гёдель не считал парадоксальные аспекты таких пространств-времен достаточным основанием для того, чтобы исключить их из списка возможных космологических моделей. По разным причинам мы сегодня, как правило, придерживаемся на этот счет более строгих взглядов, однако не всегда — см. [364]. Очень интересно бы-

[10] См. обращение к читателю в начале книги, с. 18.



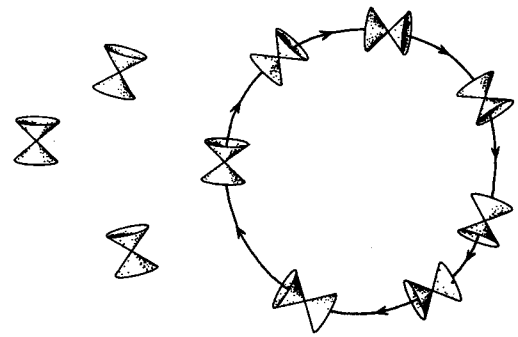


Рис. 7.15. Достаточно сильный наклон световых конусов в пространстве-времени может привести к возникновению замкнутых времениподобных линий.

ло бы увидеть реакцию Гёделя на ту роль, какую мы отведем таким пространствам-временам чуть ниже!) Хотя представляется вполне разумным исключить пространственно-временны́е геометрии с замкнутыми времениподобными линиями из числа возможных описаний *классической* Вселенной, можно привести некоторые доводы в пользу того, чтобы оставить их в качестве потенциальных кандидатов на участие в *квантовых суперпозициях*. На это, собственно, и указывал Дойч. Несмотря на то, что вклады таких геометрий в общий вектор состояния могут оказаться крайне малыми, их потенциальное присутствие производит (согласно Дойчу) поразительный эффект. Если мы обратим внимание на особенности выполнения квантовых вычислений в такой ситуации, то придем, по всей видимости, к выводу, что здесь можно выполнять и *невычислимые* операции! Это обусловлено тем, что в пространственно-временны́х геометриях с замкнутыми времениподобными линиями на вход машины Тьюринга вполне можно подать полученный ею же результат, продлив таким образом ее действие до бесконечности, буде возникнет такая необходимость, — т. е. здесь ответ на вопрос «Завершается ли данное вычисление?» действительно влияет на окончательный результат квантового вычисления. Дойч пришел к выводу, что в его схеме квантовой гравитации возможны квантовые машины с оракулом.

Насколько я смог разобраться, его аргументы с тем же успехом применимы и к машинам с оракулом боле высокого порядка.

Разумеется, многие читатели сочтут, что все это следует воспринимать с надлежащей долей здорового скептицизма. В самом деле, нет никаких реальных оснований полагать, что из такой схемы может вырасти непротиворечивая (или хотя бы правдоподобная) теория квантовой гравитации. Тем не менее, в рамках собственной системы представлений идеи логичны, а с точки зрения порождения новых идей — еще и чрезвычайно интересны; я нисколько не удивлюсь, если в ту *правильную* схему квантовой гравитации, которую мы когда-нибудь все равно найдем, попадут-таки какие-нибудь существенные фрагменты гипотезы Дойча. В моем представлении, как было особо подчеркнуто в §§ 6.10 и 6.12, для корректного объединения квантовой теории и общей теории относительности необходимо изменить сами законы квантовой теории (в соответствии с процедурой **OR**). Однако тот факт, что в подходе Дойча невычислимость — даже такая, какой, по-видимому, требует вывод $\mathcal{G}^{\alpha}$, — является свойством квантовой гравитации, я рассматриваю как ценное подтверждение возможности отыскания в конечном счете места для невычислительной активности.

В завершение отметим, что те невычислимые эффекты, на которые указывает Дойч, мы получили исключительно благодаря потенциальному наклону световых конусов, предусматриваемому общей теорией относительности Эйнштейна. Если световые конусы способны наклоняться *вообще* — пусть и на те крохотные углы, что предписывает теория Эйнштейна в обычных обстоятельствах, — то значит, они *потенциально* могут наклоняться и дальше, вплоть до возникновения замкнутых времениподобных линий. Эта потенциальная возможность играет здесь вполне контрфактуальную роль (в полном согласии с квантовой теорией) — возможность совершения действия производит эффект не менее реальный, нежели само действие!

## 7.11. Время и сознательное восприятие

Вернемся к проблеме сознания. В конце концов, именно та роль, которую играет в восприятии математической истины сознание, и увлекла нас по странной дороге в не менее странное место, где мы сейчас стоим, озираясь по сторонам. Очевидно,



впрочем, что сознание отнюдь не ограничивается одним лишь восприятием математических истин. По той дороге мы пошли только потому, что нам показалось, что она нас куда-то приведет. И я почему-то подозреваю, что многим читателям не особо нравится то «где-то», куда мы, наконец, так или иначе прибыли. Однако если теперь, с высоты новых знаний, оглянуться назад, то мы, возможно, обнаружим, что некоторые из наших старых проблем представляются нам теперь в новом свете.

Среди наиболее поразительных и непосредственных свойств сознательного восприятия особо выделяется восприятие *течения времени*. Время кажется нам настолько привычным, что мы бываем немало потрясены, обнаружив, что все наши замечательно подробные теории поведения физического мира не в состоянии (пока что) практически ничего о нем рассказать. Хуже того, то, что наиболее здравые из них таки рассказывают, находится в почти полном противоречии с тем, что говорит нам о времени наше восприятие.

Согласно общей теории относительности, «время» — это всего лишь одна из координат в описании положения пространственно-временно́го события. В пространственно-временны́х описаниях, предлагаемых нам физиками, нет ничего, что выделяло бы «время» как нечто, что «течет». В самом деле, физики довольно часто используют модели пространства-времени, в которых наряду с временны́м измерением имеется лишь *одно* пространственное измерение — в таких двухмерных пространствах-временах отличить временну́ю ось от пространственной принципиально невозможно (см. рис. 7.16). И все же никто в здравом уме не станет говорить о «течении» *пространства*! Действительно, в физических задачах, где требуется вычислить будущее состояние системы на основании настоящего ее состояния (см. § 4.2), часто рассматривают так называемые временны́е эволюции. Однако эта процедура вовсе не является обязательной, и вычисления, как правило, выполняются именно так *только потому*, что мы в данном случае строим модель (математическую) *опыта восприятия нами мира* через призму «текущего» времени (которое мы, похоже, только так и воспринимаем), — а еще потому, что нам хочется научиться предсказывать будущее[10]. Исключительно благодаря особенностям нашего восприятия, в наших вычислительных моделях мира появляются неизбежные отклонения в виде временны́х эволюций (часто, но, надо признать, не всегда),

тогда как сами физические законы таких встроенных отклонений не содержат.

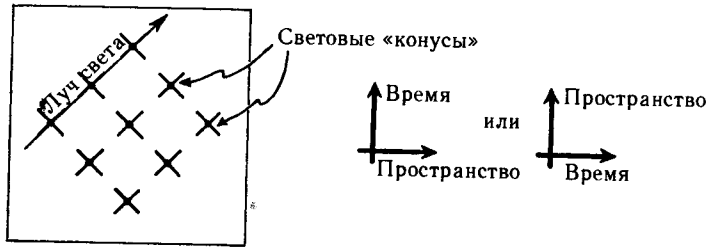



Рис. 7.16. В двухмерном пространстве-времени временна́я и пространственная оси полностью взаимозаменяемы — однако никому не приходит в голову говорить о «течении» пространства!

Более того, время для нас «течет» *только* потому, что мы обладаем сознанием. С точки зрения теории относительности, существует лишь «статическое» четырехмерное пространство-время без какого бы то ни было «течения». Пространство-время просто *есть*, и время в нем способно «течь» не больше, чем пространство. Течение времени, похоже, необходимо почему-то одному лишь сознанию, и я не удивлюсь, если отношения между сознанием и временем вдруг окажутся странными и во всем остальном.

В самом деле, было бы не совсем благоразумно чересчур тесно отождествлять феномен сознательного восприятия с его кажущимся «течением» времени и использование физиками вещественного параметра $t$ в качестве обозначения для так называемой «временно́й координаты». Во-первых, если верить теории относительности, то применительно к пространству-времени как к целому выбор параметра $t$ уникальностью не отличается. Возможны самые различные взаимно несовместные альтернативы, причем нет никаких оснований отдать предпочтение какой-то одной из них. Во-вторых, очевидно, что точная концепция «вещественного числа» имеет весьма малое отношение к сознательному восприятию нами течения времени, хотя бы по одной той причине, что мы не можем воспринимать очень малые временны́е



промежутки — скажем, порядка сотой доли секунды, не говоря уже о меньших, — тогда как физики способны работать и с временны́ми масштабами порядка $10^{-25}$ с (что с успехом демонстрирует точность квантовой электродинамики, т. е. квантовой теории взаимодействия электромагнитных полей с электронами и другими заряженными частицами) или, возможно, еще меньшими, вплоть до планковского времени $10^{-43}$ с. Более того, согласно математической концепции времени, выраженного в виде вещественного числа, предела малости, после достижения которого концепция должна потерять всякий смысл, *нет вообще* — вне зависимости от того, имеет эта концепция физический смысл во всех масштабах величин или нет.

Возможно ли сказать что-либо более конкретное о взаимоотношениях между сознательно воспринимаемым временем и параметром $t$, который физики называют «временем» и используют в таком качестве в своих физических описаниях? Можно ли каким-либо образом экспериментально установить, «когда» именно, по отношению к этому физическому параметру, «на самом деле» происходит субъективное восприятие? Имеет ли какой-нибудь *объективный смысл* высказывание о том, что то или иное осознаваемое событие происходит в тот или иной момент времени? По правде говоря, кое-какие эксперименты, имеющие определенное отношение к данной проблеме, действительно проводились, однако результаты их оказались весьма неоднозначными, а следствия из этих результатов — почти парадоксальными. Описание отдельных экспериментов я приводил в НРК, с. 439–444, однако, думаю, будет уместно рассмотреть их здесь снова.

В середине 1970-х годов Г. Г. Корнхубер с коллегами (см. [78]), используя метод электроэнцефалограммы (ЭЭГ), записали электрические сигналы в различных точках на головах нескольких добровольцев с целью установить возможные временны́е соответствия между электрической активностью мозга и актами проявления *свободы воли* (*активного* аспекта сознания). Испытуемых просили сгибать указательный палец через различные промежутки времени, причем момент сгибания пальца *полностью определял* сам доброволец; тем самым экспериментаторы надеялись проследить связь между активностью мозга, направленной на осуществление «волевого акт» сгибания пальца, с собственно движением. Для получения сколько-нибудь достоверной информации с датчиков ЭЭГ каждый опыт повторяли по несколько раз,

а затем полученные данные усредняли. Результат оказался весьма удивительным: *прежде* чем испытуемый сгибал палец, записанный электрический потенциал постепенно нарастал в течение некоторого времени (от секунды до полутора секунд). Означает ли это, что между сознательным волевым актом и обусловленным им действием должна пройти целая секунда или даже больше? Насколько осознавали сами испытуемые, между решением согнуть палец и его действительным сгибанием проходило лишь краткое мгновение — никак не секунда, и уж конечно же, не больше. (Заметим, что «запрограммированное» время реакции на внешний стимул гораздо меньше и составляет приблизительно пятую долю секунды.)

Отсюда можно, по-видимому, заключить, что *либо* (i) сознательный акт «свободной воли» есть чистая иллюзия, поскольку он, в некотором смысле, заранее запрограммирован предшествующей бессознательной активностью мозга, *либо* (ii) воле, возможно, отведена роль «на последнюю минуту», т. е. она может иногда (но не всегда) отменить действие, которое бессознательно готовилось в течение последней секунды, *либо* (iii) субъект на самом деле пожелал согнуть палец на секунду (или больше) раньше, чем палец согнулся, однако ошибочно воспринимает (непротиворечивым образом) это так, будто сознательный акт произошел в значительно более поздний момент времени, непосредственно перед тем, как палец действительно был согнут.

Позднее Бенджамин Либет (с группой сотрудников) повторил эксперимент Корнхубера, но с некоторыми модификациями, направленными на уточнение момента времени, в который происходит волевой акт, направленный на сгибание пальца: испытуемому было предложено отмечать положение стрелки часов в момент принятия решения (см. [238, 239]). Новый эксперимент в целом подтвердил полученные ранее выводы, за исключением вывода (iii); сам Либет, похоже, склонялся к (ii).

В других экспериментах Либет и Файнстайн [240] исследовали временны́е соответствия *сенсорных* (или *пассивных*) аспектов сознания. Испытуемыми являлись добровольцы, давшие согласие на помещение электродов в область мозга, связанную с приемом сенсорных сигналов от определенных участков кожи. Наряду с прямой стимуляцией электродами, время от времени стимулировался и соответствующий участок кожи. Общий результат эксперимента таков: прежде чем испытуемые осознавали, что они что-то ощущают, проходило приблизительно полсекунды нейронной активности (с некоторыми вариациями в зависимости от обстоятельств), хотя у них создалось впечатление, что при прямой стимуляции они узнаю́т о возникновении ощущения раньше, чем при реальной стимуляции кожи.



Каждый из этих экспериментов сам по себе парадоксальным не является, разве что внушает некоторое беспокойство. Возможно, кажущиеся сознательными решения и в самом деле принимаются на *бессознательном* уровне, причем раньше (по меньшей мере, на секунду). Возможно, и в самом деле *необходимо* полсекунды активности мозга, прежде чем мы действительно осознаём то, что ощущаем. Однако если свести эти два вывода вместе, то получается, что в любом действии, где внешний стимул вызывает сознательно обусловленную реакцию, эта самая реакция возникает с запаздыванием, составляющим от секунды до полутора. Пока не пройдет полсекунды, не произойдет осознания; а если мы решим это осознание применить к делу, то нам придется запустить неторопливую машину свободной воли, что, возможно, задержит реакцию еще на секунду.

Неужели наши сознательные реакции действительно настолько медлительны? В обычном разговоре, например, такая задержка почему-то не наблюдается. Если принять вывод (ii), то получается, что большая часть актов реакции полностью бессознательна, хотя время от времени человек оказывается способен отменить эту реакцию, заменив ее (где-то через секунду) сознательным волевым актом. Однако если реакция обычно *бес*сознательна, то у сознания (если, конечно, оно не сравнится с ней по скорости) нет ни одного шанса успеть ее отменить — когда начинает действовать сознательный волевой акт, бессознательная реакция уже давно запущена, и предпринимать что-либо слишком поздно! Таким образом, либо сознательные акты могут *иногда* действовать быстро, либо бессознательная реакция и *сама* на секунду запаздывает. В этой связи вспомним, что «запрограммированная» бессознательная реакция может произойти гораздо быстрее — через пятую долю секунды или около того.

Разумеется, быстрая (скажем, за пятую долю секунды) бессознательная реакция все еще возможна, если мы принимаем вывод (i), согласно которому система бессознательных реакций полностью игнорирует любые возможные попытки позднейшей сознательной (сенсорной) активности. В этом случае (а ситуация

с выводом (iii), поверьте, еще хуже) сознание в достаточно быстром разговоре способно выступать единственно в роли зрителя, сознательно воспринимающего нечто вроде «записи» давно прошедшего спектакля.

Здесь в действительности нет никакого противоречия. Вполне возможно, что эволюция произвела на свет сознание как раз для неторопливых размышлений, и очевидно, что в ситуации, требующей сколько-нибудь быстрых действий, сознание оказывается не более чем пассажиром. Вся первая часть книги, если помните, посвящена именно такому сознательному созерцанию (математическому пониманию), которое и впрямь славится своей медлительностью. Может быть, способность к сознательному восприятию развилась у нас исключительно ради вот такой вот неспешной созерцательной мыслительной деятельности, тогда как более быстрые по времени реакции полностью бессознательны по своему происхождению — хотя и сопровождаются запаздывающим сознательным восприятием, не играющим, впрочем никакой активной роли.

Все это, конечно же, правильно — сознание действительно «берет свое», когда располагает достаточным временем для работы. Однако должен признать, что я не верю, что сознание может не играть абсолютно *никакой* роли в умеренно быстрой деятельности, такой как обычный разговор — или настольный теннис, футбол и гонки на мотоциклах, если уж на то пошло. Мне представляется, что в логике предыдущих рассуждений имеется одна фундаментальная дыра, и в роли этой дыры выступает допущение об осмысленности точного хронометража сознательных событий. Можно ли вообще говорить о каком-то реальном «моменте времени», в котором происходит акт сознательного восприятия, предполагая к тому же, что этот самый «момент восприятия» должен непременно предшествовать моменту проявления того или иного эффекта «реакции свободной воли» на упомянутый акт восприятия. Учитывая те аномальные взаимоотношения между сознанием и собственно физической природой времени, что описаны в начале этого параграфа, я полагаю (по меньшей мере) возможным, что *никакого* выраженного «момента времени», в котором происходит акт сознательного восприятия, в природе не существует[11].

Самой умеренной из всех возможностей в свете вышесказанного представляется нелокальный разброс во времени, при-



дающий связи сознательного восприятия с физическим временем некоторую неустранимую размытость. Однако я подозреваю, что тут работает нечто гораздо более тонкое и непонятное. Если сознание является феноменом, который невозможно понять на физическом уровне без существенного привлечения квантовой теории, то вполне может оказаться так, что **Z**-загадки этой самой теории входят в противоречие с нашими — такими на вид безупречными! — умозаключениями относительно причинности, нелокальности и контрфактуальности, которые, возможно, и впрямь свойственны отношениям между сознанием и свободной волей. Например, какую-то роль, возможно, играет та контрфактуальность, какую мы наблюдали в задаче об испытании бомб (см. §§ 5.2 и 5.9): на поведение может повлиять один лишь факт *возможности* некоего действия или мысли, даже если в действительности никто ничего не сделал и не подумал. (Это может лишить всякой силы некоторые кажущиеся вполне логичными заключения — скажем, то, с помощью которого мы исключаем возможность правильности вывода (ii).)

В общем и целом, ко всем логичным на первый взгляд выводам касательно упорядочивания событий во времени в присутствии квантовых эффектов следует подходить очень осторожно (что будет особо подчеркнуто в следующем параграфе, где мы рассмотрим проблему с точки зрения ЭПР-феноменов). И напротив, *если*, в том или ином проявлении сознания, классические рассуждения о расположении событий во времени приводят нас к явно противоречивому заключению, то это совершенно недвусмысленно указывает на присутствие квантовых процессов!

## 7.12. ЭПР-феномены и время: необходимость в новом мировоззрении

Есть основания относиться к нашему физическому представлению о времени с некоторой подозрительностью — причем не только в отношении сознания, но и в отношении собственно физики, когда в дело вступают квантовые нелокальность и контрфактуальность. Если придерживаться строго «реалистичного» взгляда на вектор состояния $|\psi\rangle$ в ситуациях ЭПР-типа (см. §§ 6.3 и 6.5, где живописуются трудности, подстерегающие тех, кто этого *не* делает), то перед нами в полный рост встает

фундаментально головоломная проблема. Проблемы такого рода вырастают в труднопреодолимые препятствия при разработке, например, детальной ГРВ-теории (см. §6.9) или любой другой подобной теории, затрагивая потенциально и любую схему **OR**-типа, вроде той, что я предлагаю в §6.12.

Вспомним магические додекаэдры из §5.3 и объяснение их поведения, представленное в §5.18, и спросим себя, какая из двух следующих возможностей отражает «реальное» положение дел. Может быть, именно нажатие на кнопку на додекаэдре моего *коллеги* вызывает мгновенную редукцию (и расцепление) исходного сцепленного совокупного состояния — т. е. по нажатии *его* кнопки атом в моем додекаэдре мгновенно переходит в новое, расцепленное состояние, и именно *это* редуцированное состояние и отменяет все остальные варианты развития событий, которые могли бы реализоваться после моего *более позднего* нажатия на кнопку? Или, может быть, это *я* нажимаю на кнопку первым, воздействуя на исходное сцепленное состояние, результатом чего становится мгновенная редукция состояния атома в додекаэдре моего коллеги, и теперь уже он не может ничего поделать, на какие бы кнопки он ни нажимал? Для получаемого результата совершенно неважно, какой вариант рассмотрения проблемы мы выберем (о чем мы уже говорили в §6.5). И хорошо, что неважно, потому что если бы было важно, то мы получили бы нарушение принципов эйнштейновской теории относительности, согласно которой «одновременность» в случае удаленных (пространственноподобно разделенных) событий не может иметь никаких наблюдаемых эффектов. Однако если мы полагаем, что вектор $|\psi\rangle$ есть отражение *реальности*, то реальность эта в двух представленных картинах получается различной. Кто-то, возможно, сочтет это расхождение достаточной причиной для того, чтобы отказаться от такого «реалистичного» взгляда на $|\psi\rangle$. Другие же, напротив, отыщут иные строгие доводы в пользу реальности $|\psi\rangle$ (см. §6.3) — и приготовятся вышвырнуть эйнштейновскую картину мира за борт.

Я склоняюсь к тому, чтобы попытаться примирить обе эти точки зрения — квантовый реализм и дух релятивистского пространства-времени. Однако для этого потребуется фундаментальный пересмотр наших современных представлений о физической реальности. Вместо того, чтобы настаивать на том, что способ описания квантового состояния (или даже пространства-времени) непременно должен следовать из привычных описаний, мы должны отыскать нечто совершенно иное, хотя и эквивалентное математически (по крайней мере, на первых порах) этим самым описаниям.



Более того, имеется и хороший прецедент. Прежде чем Эйнштейн пришел к общей теории относительности, нас полностью устраивала уютная и замечательно точная ньютоновская теория гравитации, согласно которой движущиеся в плоском пространстве частицы притягивали друг друга в соответствии с обратно-квадратичным законом всемирного тяготения. Внесение каких-то фундаментальных изменений в такую гармоничную картину непременно разрушило бы великолепную точность ньютоновской схемы. И тем не менее, именно такое фундаментальное изменение Эйнштейн и предложил. Его альтернативный взгляд на гравитационную динамику полностью переписал прежнюю картину. Пространство больше не является плоским (и вообще, это уже даже не «пространство», а «пространство-время»), а гравитационных сил в природе не существует — есть приливные эффекты искривлений пространства-времени. Что касается частиц, то они, как выясняется, и не движутся вовсе, будучи представлены «статическими» кривыми на пространстве-времени. Разрушило ли все это замечательную точность теории Ньютона? Ни в малейшей степени; теория стала еще точнее, хотя, казалось бы, уже и некуда! (См. §4.5.)

Можно ли ожидать, что нечто подобное произойдет и с квантовой теорией? Думаю, что вероятность такого исхода крайне высока. Просто для этого необходимо фундаментальное изменение *мировоззрения*, поэтому представить себе сейчас умозрительно природу предстоящего изменения чрезвычайно трудно. Более того, оно несомненно будет выглядеть, как самый настоящий бред!

В заключение я хочу рассказать о двух таких бредовых идеях — ни одна из них, к сожалению, не достигает необходимой степени бредовости, однако у каждой имеются свои достоинства. Первую предложили Якир Ахаронов и Лев Вайдман [2] (а также Коста де Борегар [61] и Пол Вербос [381]). Суть идеи в том, что квантовая реальность описывается *двумя* векторами состояния, один из которых направлен во времени вперед от последней редукции **R** (нормальное направление), а другой — *назад*, от сле-

дующей редукции **R** в будущем. Второй вектор состояния[11] ведет себя «телеологически» — он обусловлен тем, чему предстоит случиться с ним в будущем, а не тем, что с ним уже произошло в прошлом; многие, боюсь, сочтут это его свойство неприемлемым. Однако результаты эта модификация дает в точности те же, что и стандартная квантовая теория, поэтому исключить новую теорию только на этом основании не удастся. Ее *преимущество* перед стандартной квантовой теорией заключается в том, что она позволяет получить полностью объективное описание состояния в ЭПР-ситуациях, которые теперь можно рассматривать в терминах пространства-времени сообразно духу эйнштейновской теории относительности. Таким образом, новая теория предлагает решение (пусть и своеобразное) головоломной проблемы, о которой мы упоминали в начале этого параграфа, — однако лишь за счет введения квантового состояния, отличающегося телеологическим поведением, что не всем по душе. (Лично я нахожу эти телеологические аспекты вполне приемлемыми, коль скоро они не вступают в конфликт с действительным физическим поведением.) За подробностями отсылаю читателя к соответствующей литературе.

Другая идея, о которой я хотел упомянуть, — это *теория твисторов* (см. 7.17). Поводом для создания этой теории послужили все те же ЭПР-головоломки, однако решения для них она (как таковая) *пока* не предоставляет. Ее сила в другом — в неожиданных и изящных математических описаниях некоторых фундаментальных физических концепций (таких, например, как электромагнитные уравнения Максвелла, см. §4.4 и НРК, с. 184–187, приобретающие в теории твисторов привлекательную математическую формулировку). Имеется и нелокальное описание пространства-времени, где каждый луч света представляется в виде точки. Именно эта пространственно-временна́я нелокальность и связывает теорию твисторов с квантовой нелокальностью ЭПР-ситуаций. Кроме того, в основе теории лежат *комплексные числа* и соответствующая геометрия, чем достигается тесная взаимосвязь между комплексными коэффициентами **U**-квантовой теории и структурой пространства-времени. В частности, фундаментальную роль приобретает сфера Римана (см. §5.10), связанная здесь со световым конусом пространственно-временно́й точки (а также с «небесной сферой» находящегося в этой точке наблюдателя). (Неформальное описание идей, имеющих отношение к данной теме, приводится в книге Дэвида Пита [287]; относительно краткое, но строгое описание теории твисторов можно найти в работе Стивена Хаггета и Пола Тода [209][(12)].)

[11] Есть некий математический смысл в том, что эволюционирующий в обратном направлении вектор состояния обозначается как «бра-вектор», $\langle\phi|$, тогда как вектор, эволюционирующий нормально, получает стандартное обозначение «кет-вектора», $|\psi\rangle$. Такую пару векторов состояний можно рассматривать как произведение $|\psi\rangle\langle\phi|$. Это обозначение фигурирует также в формализме матриц плотности из §6.4.



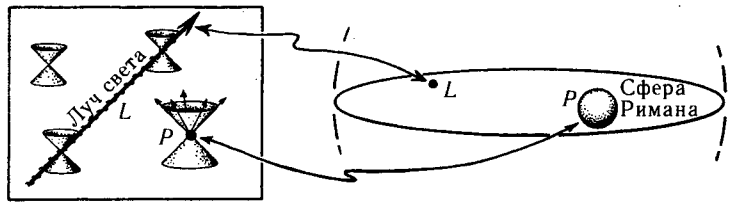



Рис. 7.17. Теория твисторов предлагает альтернативную физическую картину пространства-времени, где лучи света представлены точками, а события — целыми сферами Римана.

Думаю, продолжать углубляться в эти идеи дальше будет не совсем уместно. Я упомянул о них только для того, чтобы показать, что существует множество возможностей изменить нашу уже и так чрезвычайно точную картину физического мира, превратить ее в нечто, совершенно отличное от того, к чему мы успели привыкнуть за прошедшие десятилетия. Такое изменение должно удовлетворять требованию совместимости — иначе говоря, с помощью нового описания мы должны суметь воспроизвести все успешные результаты **U**-квантовой теории (равно как и общей теории относительности). Однако оно должно также позволить нам продвинуться за сегодняшние пределы и осуществить физически корректную модификацию квантовой теории с целью замены процедуры **R** на какой-либо реальный физический процесс. В этом (по меньшей мере) я убежден твердо; мне также представляется, что такая «корректная модификация» будет включать в себя некую **OR**-подобную процедуру, основан-

ную на идеях, изложенных в §6.12. Напомню, что теории, сочетающие в себе относительность с «реалистичной» редукцией состояний (такие как ГРВ-теория) сталкиваются сегодня с труднопреодолимыми проблемами (в частности, связанными с сохранением энергии). Это лишь укрепляет мою собственную уверенность: прежде чем мы сможем хоть сколько-нибудь серьезно продвинуться в понимании фундаментальных вопросов физики, мы должны фундаментально изменить наши представления о мире.

Нисколько не сомневаюсь я и в том, что истинный прогресс в физическом понимании феномена *сознания* попросту невозможен без все того же фундаментального изменения в нашем физическом мировоззрении.

## Примечания

1. См., напр., [242].
2. См. [128], [139], [11] и [134].
3. Напр., [101].
4. См. [184], [183] и [186]. В недавней работе [371] указывается, что такая обработка информации может осуществляться только в микротрубочках, организованных в виде так называемых «А-решеток» (именно эта структура и показана на рис. 7.4, 7.8 и 7.9), тогда как более распространенная организация в виде «В-решетки» (с характерным «швом», проходящим вдоль трубки, см. [254]), для обработки информации не годится.
5. См. [229] (доступно о клатринах) и [66] (популярное описание фуллеренов).
6. См. [363].
7. Например, полученное Хамероффом время переключения димеров тубулина, по-видимому, согласуется с частотой, предсказанной Фрёлихом ($\sim 5 \times 10^{10}$ Гц).
8. См., напр., [211, 212] и [348, 349].
9. Эта идея описана в одном из черновых вариантов статьи Дэвида Дойча «Квантовая механика вблизи замкнутых времениподобных линий» [85], однако в опубликованную статью она не попала. Дэвид уверил меня в том, что он убрал этот кусок из окончательного варианта статьи не потому, что счел идею «ошибочной», а потому лишь, что она не имела непосредственного отношения к теме статьи. Как



бы то ни было, в рамках моей собственной «темы» ценность идеи заключается не в том, чтобы она была «корректной» по меркам той или иной системы взглядов на квантовую гравитацию — поскольку такой системы взглядов (непротиворечивой) в настоящий момент все равно нет, — но в том, чтобы она содержала в себе потенциал для дальнейших исследований, а этого в идее Дойча с избытком!

10. Во всяком случае, в рамках наших обычных физических представлений о времени «течение» времени в будущее ничем не отличается от «течения» времени в прошлое. (Однако, благодаря второму закону термодинамики, осуществить эффективное «*после*сказание» прошлого с помощью временно́й эволюции уравнений динамики невозможно.)
11. См. также [81].

    У людей, видевших фильм «Краткая история времени», в котором рассказывается о Стивене Хокинге и его работе, могло создаться весьма занятное представление о моих взглядах на связь сознания с течением времени. Пользуясь предоставившейся возможностью, заявляю, что все это — чистое недоразумение, вызванное ошибками при монтаже фильма.
12. Для получения более подробных сведений о твисторах см. также [302], [378] и [16].

# 8

# ВОЗМОЖНЫЕ ПОСЛЕДСТВИЯ

## 8.1. Искусственные разумные «устройства»

Какие же выводы должны мы сделать, исходя из предыдущих рассуждений, о предельном потенциале искусственного интеллекта? В первой части книги было недвусмысленно показано, что никакое развитие технологий производства электронных роботов с компьютерным управлением не приведет в конечном итоге к созданию *действительно* разумной искусственной машины — в том смысле, что машина будет способна понимать, что она делает, и действовать на основе этого понимания. Электронные компьютеры, несомненно, играют очень важную роль в прояснении многих вопросов, связанных с ментальными феноменами (возможно, прежде всего тем, что́ наглядно показывают, что подлинными ментальными феноменами *не* является), не говоря уже об их чрезвычайной полезности и бесценном вкладе в научный, технический и социальный прогресс. Вывод, впрочем, однозначен: компьютеры делают что-то принципиально отличное от того, что делаем *мы*, сосредоточивая сознательное внимание на очередной проблеме.

Однако, как можно было понять из продолжения нашего разговора во второй части, я ни в коем случае не утверждаю, что создать подлинно разумное *устройство* совершенно невозможно; просто такое устройство не будет являться «машиной» — в том конкретном смысле, что «машиной» управляет компьютер. В основе его работы должны будут лежать те же физические процессы, которые ответственны за возникновение нашего собственного сознания. Поскольку физической теории таких процессов в нашем распоряжении еще нет, представляется несколько преж-



девременным делать какие-то умозаключения относительно того, будет ли вообще построено такое устройство, и если будет, то когда. Тем не менее, в рамках поддерживаемой мною точки зрения $\mathscr{C}$ (см. § 1.3), согласно которой мышление может быть в конечном счете объяснено научно, хотя и с привлечением понятия невычислимости, создание этого устройства вполне допускается.

Не думаю, что такое устройство непременно должно быть по своей природе биологическим. Более того, я не думаю, что между биологией и физикой (или между биологией, химией и физикой) проходит какая-то принципиально непреодолимая граница. Биологическим системам действительно зачастую присуща тонкость и сложность организации, далеко превосходящая даже наиболее изощренные из наших (порой очень и очень изощренных) физических построений. Однако совершенно очевидно, что мы все еще находимся на очень раннем этапе физического понимания нашей Вселенной — в особенности, феноменов, имеющих отношение к мышлению. Таким образом, можно ожидать, что в будущем сложность наших физических построений существенно возрастет. Можно предположить, что в этом будущем усложнении немалую роль сыграют физические эффекты, о которых мы сегодня имеем весьма смутное представление.

Не вижу причин сомневаться в том, что в не столь отдаленном будущем некоторые из приводящих нас сейчас в недоумение эффектов (**Z**-загадок) квантовой теории найдут удивительные применения в самых разнообразных областях. Уже сегодня предлагаются идеи использования квантовых эффектов в криптографии, что позволяет достичь результатов, недоступных классическим устройствам. В частности, имеются теоретические разработки, предполагающие существенное использование квантовых эффектов (см. [26]) и направленные на отыскание способа передачи секретной информации от отправителя к получателю таким образом, чтобы перехват сообщения третьей стороной был невозможен без обнаружения факта перехвата. На основе этих идей уже были разработаны экспериментальные устройства, которые, несомненно, найдут через несколько лет самое широкое коммерческое применение. В области криптографии было предложено и множество других схем, так или иначе использующих квантовые эффекты, и можно сказать, что вчера еще не существовавшая наука *квантовая криптография* сегодня развивается бурными темпами. Более того, возможно, что когда-нибудь

мы действительно сможем построить *квантовый компьютер*, однако на данный момент соответствующие теоретические разработки еще весьма далеки от практической реализации, и пока весьма сложно предсказать, когда мы увидим (и увидим ли вообще) их физическое воплощение (см. [277, 278]).

Еще сложнее предсказать возможность (и время) создания устройства, работа которого описывается физической теорией, нам еще даже не *известной*. Я утверждаю, что такая теория необходима для понимания физики, лежащей в основе устройства, функционирующего невычислимым образом; под «невычислимым» здесь понимается «недоступным для машины Тьюринга». Согласно приведенной выше аргументации, прежде чем рассматривать саму возможность создания такого устройства, мы должны отыскать надлежащую физическую теорию редукции квантового состояния (**OR**) — а насколько мы сейчас далеки от такой теории, сказать очень сложно. Возможно также, что возникнут какие-то дополнительные неожиданные трудности, обусловленные неизвестными пока специфическими особенностями будущей **OR**-теории.

Как бы то ни было, если мы хотим построить такое невычислительное устройство, нам все равно придется, *я думаю*, начать с отыскания теории. Впрочем, возможно, что и не придется: история помнит немало случаев, когда между открытием новых необычных физических эффектов и их теоретическим объяснением проходило много лет. Хорошим примером может послужить сверхпроводимость, обнаруженная экспериментально (Хейке Камерлинг-Оннесом в 1911 году) почти за пятьдесят лет до того, как Бардин, Купер и Шриффер получили наконец (в 1957 году) полное квантово теоретическое ей объяснение. В 1986 году была открыта высокотемпературная сверхпроводимость (см. [343]) — также при полном отсутствии предварительных чисто теоретических оснований верить в ее существование. (По состоянию на начало 1994 года адекватного теоретического объяснения этому феномену у нас все еще нет.) С другой стороны, если речь идет о невычислимых процессах, неясно даже, каким образом вообще можно *определить*, что поведение данного неодушевленного объекта является невычислимым. Вся концепция вычислимости опирается в значительной степени на *теорию*, и непосредственное наблюдение в этом случае мало что дает. Однако в рамках той или иной невычислительной теории вполне может



быть описано поведение, которое демонстрирует невычислимые аспекты этой самой теории и которое вполне можно исследовать экспериментально и регистрировать с помощью каких-то реальных приборов. Я подозреваю, что в отсутствие теории вероятность наблюдать или регистрировать невычислимое поведение в каких-либо физических объектах исключительно мала.

А теперь давайте попробуем вообразить, что требуемая физическая теория — т. е., как я показал выше, невычислительная **OR**-теория редукции квантового состояния — у нас уже *есть*; кроме того, мы располагаем и некоторыми экспериментальными подтверждениями этой теории. Что нам нужно сделать для того, чтобы создать *разумное* искусственное устройство? А *ничего* — располагая одной лишь этой теорией, мы ничего сделать не сможем. Понадобится еще один теоретический прорыв — тот, что объяснит нам, как именно соответствующая организация, задействуя надлежащим образом невычислимые **OR**-эффекты, порождает сознание. Я, например, не имею ни малейшего понятия, что это может оказаться за теория. Как и в упомянутых выше примерах со сверхпроводимостью, есть вероятность, что на устройство с требуемыми свойствами кто-нибудь наткнется до некоторой степени случайно раньше, чем будет разработана корректная теория сознания. Само собой разумеется, вероятность эта крайне ничтожна — разве что воспользоваться неким дарвиновским эволюционным процессом, т. е. предположить, что разум возникнет сам собой, просто по причине непосредственных преимуществ, которые обладание разумом дает его обладателю, задолго до того, как этот самый обладатель сможет понять, каким же образом все произошло (как, собственно говоря, и случилось когда-то с нами!). Процесс этот, безусловно, будет чрезвычайно длительным, особенно если вспомнить, сколько времени потребовалось нашему с вами разуму для проявления себя в качестве такового. Возможно, гораздо более удовлетворительным путем к созданию искусственного разумного устройства покажется читателю прямое заимствование тех на первый взгляд беспорядочных, но все же замечательно эффективных и уместных процедур, которыми мы сами многие тысячелетия с успехом пользуемся.

Разумеется, ничто из вышесказанного отнюдь не отменяет нашего желания узнать, что́ же все-таки происходит там, в глубинах сознания, что́ делает разум разумом. Я и сам хочу это узнать. Все, о чем я говорил в этой книге, является, в сущности,

доказательством одного простого утверждения: то, что происходит в сознании, отнюдь *не* сводится к совокупности исключительно вычислительных процессов — как многие сегодня полагают — и не может быть *в полной мере* понято до тех пор, пока мы не достигнем более глубокого понимания природы материи, времени, пространства и тех законов, что ими управляют. Нам потребуются также гораздо более обширные и подробные знания в области физиологии мозга, особенно на микроскопических уровнях, избегавших до недавних пор внимания исследователей. Мы должны больше узнать об условиях, при которых сознание возникает и исчезает, о его любопытных отношениях с временем, о применениях сознания и о преимуществах обладания им — и о многих других вещах, допускающих объективное исследование. Таким образом, перед нами открывается широчайшее поле деятельности, обещающее несомненный прогресс в самых разных областях.

## 8.2. Что компьютеры умеют делать хорошо... и что не очень

Даже зная о том, что существующая концепция компьютера *не позволяет* достичь ни подлинной разумности, ни какого бы то ни было осознания себя, ни в коем случае не следует сбрасывать со счетов огромную мощь современных компьютеров, которая в ближайшей перспективе, по-видимому, увеличится и вовсе до невообразимых пределов (см. §§ 1.2, 1.10 и [267]). Пусть эти машины и не *понимают* того, что они делают, они делают это невероятно быстро и точно. Не смогут ли компьютеры таким образом (пусть и неразумным) достичь — к тому же с большей эффективностью — тех же результатов, для получения которых мы используем разум? Можем ли мы сказать заранее, в каких областях компьютерные системы добьются больших успехов, а в каких им никогда не удастся превзойти разум?

Уже сегодня компьютеры замечательно играют в шахматы — приближаясь к уровню лучших гроссмейстеров-людей. В шашки компьютер «Чинук» обыграл всех противников за исключением абсолютного чемпиона мира Мариона Тинсли. А вот в древней восточной игре го компьютеры, как выясняется, не сильны. Компьютер здесь получает преимущество только в том случае, когда продолжительность хода ограничена; если же дать



человеку достаточно времени на ход, то компьютер, как правило, оказывается в проигрыше. Шахматные задачи глубиной в два-три хода компьютер решает практически мгновенно, вне зависимости от того, насколько сложной находит задачу человек. С другой стороны, простая по замыслу, но требующая для решения, скажем, пятьдесят или сто ходов задача может привести к полному поражению компьютера, тогда как опытный шахматист-человек, возможно, никаких трудностей и не встретит (см. также § 1.15 и рис. 1.7).

Эти особенности по большей части объясняются различиями в способностях, присущих компьютеру и человеку. Компьютер всего лишь выполняет вычисления, не понимая при этом, что он делает, — хотя он и пользуется опосредованно тем пониманием, которое *программисты* вложили в написание программы. Компьютер может хранить и использовать большой объем информации; человек, впрочем, на это тоже способен. Компьютер может многократно, чрезвычайно быстро и точно выполнять предписанные ему программистами операции; его действия абсолютно бездумны, но по скорости и точности далеко превосходят возможности любого человека. Игрок-человек оценивает ситуацию и составляет осмысленные планы, располагая при этом общим пониманием игры вообще и данной конкретной позиции в частности. Эти способности компьютеру абсолютно недоступны, однако недостаток действительного понимания он зачастую с успехом заменяет вычислительной мощью.

Предположим, что количество возможных вариантов, которые компьютеру необходимо рассмотреть за один ход, равно, в среднем, $p$; тогда при глубине в $m$ ходов компьютеру придется рассмотреть $p^m$ альтернатив. Если расчет каждой альтернативы занимает в среднем время $t$, то полное время $T$, необходимое для расчета задачи на такую глубину, составит

$$T = t \times p^m.$$

В шашках число $p$ не бывает очень большим — скажем, четыре, — что позволяет компьютеру за отведенное ему время просчитывать дальнейшую игру на значительную глубину, вплоть до двадцати ходов ($m = 20$), тогда как в игре го нередки ситуации, когда $p = 200$, и сравнимая по мощности компьютерная система справится в этом случае не более чем с пятью ($m = 5$) ходами или около того. Шахматы располагаются где-то посередине. Кроме

того, необходимо учесть, что человеческие оценки и понимание гораздо медленнее, нежели компьютерные вычисления (для человека $t$ велико, для компьютера — мало́), однако с помощью этих оценок человек способен значительно сократить эффективное число $p$ (для человека эффективное значение $p$ мало́, для компьютера — велико), поскольку достойной дальнейшего рассмотрения человек сочтет лишь малую часть всех доступных альтернатив.

В общем случае из этого следует, что в играх, где $p$ велико, но может быть значительно уменьшено посредством понимания и оценки, относительное преимущество получает игрок-человек. При достаточно большом $T$ человеческая способность сократить «эффективное $p$» увеличивает $m$ в формуле $T = t \times p^m$ гораздо быстрее, нежели этого можно добиться, уменьшая $t$ (что как раз очень хорошо умеют делать компьютеры). Однако при *малом* полном времени $T$ более эффективным оказывается уменьшение $t$ (поскольку существенные для данной игры значения $m$ будут, скорее всего, тоже небольшими). Эти выводы представляют собой простые следствия из «экспоненциальной» формы выражения $T = t \times p^m$.

Приведенное рассуждение страдает некоторой упрощенностью, однако суть его, полагаю, достаточно ясна. (Если вы не математик, но хотите получить представление о том, как ведет себя выражение $t \times p^m$, попробуйте подставить в него различные значения $t$, $p$ и $m$.) Я не вижу особого смысла углубляться здесь в подробности, но кое-что, думаю, следует прояснить. Кто-то, возможно, полагает, что «большая глубина вычисления», выражаемая числом $m$, — это вовсе не то, чего стремится достичь игрок-человек. Спешу разуверить: *в действительности* человек стремится именно к этому. Когда игрок-человек определяет ценность позиции на несколько ходов вперед, а затем решает, что дальше ее просчитывать смысла нет, такое вычисление является *в действительности* вычислением гораздо большей глубины, поскольку человеческая оценка охватывает и возможный эффект нескольких последующих ходов. Как бы то ни было, с помощью упрощенных соображений такого рода можно в общих чертах понять, почему научить компьютер хорошо играть в го гораздо сложнее, чем научить его хорошо играть в шашки, почему у компьютеров лучше получается решать короткие шахматные задачи и почему компьютеры получают относительное преимущество в играх с ограничением на время хода.



Подчеркнем еще раз главное отличие: человеческий мозг обладает способностью, какой компьютер принципиально лишен, — мозг способен выносить *суждения*, основанные на *понимании*. Именно это различие и приводит к следствиям, описанным в общем виде в вышеприведенных простых рассуждениях (а также в рассуждениях относительно шахматной задачи, представленной на рис. 1.7 в § 1.15). Сознательное понимание — процесс сравнительно медленный, однако он может значительно сократить число альтернатив, требующих серьезного рассмотрения, существенно увеличив таким образом *эффективную* глубину вычисления. (По достижении определенной точки необходимость в рассмотрении отдельных альтернативных вариантов и вовсе отпадает.) И вообще, всем, кому интересно, чего компьютеры могут достичь в будущем, я, думается, могу дать хорошую подсказку: попытайтесь ответить на вопрос, требуется ли для выполнения той или иной задачи подлинное понимание. Многие вещи в нашей повседневной жизни не требуют для своего выполнения какого-то особого понимания, и вполне возможно, что с ними отлично справятся роботы с компьютерным управлением. Уже сейчас существуют управляемые искусственными нейронными сетями машины, успешно выполняющие такого рода задачи. Например, машины научились достаточно хорошо распознавать лица, производить геологическую разведку, находить по звуку неполадки в работе различных механизмов, разоблачать мошенничества с кредитными картами и т. д.[1] Там, где применение таких машин возможно, их эффективность в общем случае приближается к средней эффективности экспертов-людей (а порой и превосходит ее). Однако вследствие особенностей необходимого в данном случае «восходящего» программирования, мы не увидим здесь того уровня мощной машинной «компетентности», какой присущ нисходящим системам (скажем, шахматным компьютерам), или того, что — еще более впечатляюще — демонстрируют компьютеры при выполнении обыкновенных численных расчетов, в каковой области даже лучшие вычислители-люди и близко не подходят к производительности средних по сегодняшним меркам компьютеров. Что же касается задач, с которыми эффективно справляются искусственные нейронные сети (восходящего типа), то задействуемое в выполнении таких задач *людьми* понимание, если честно, едва ли превышает способности компьютеров, поэтому в таких областях от компьютеров можно ожидать некото-

рого ограниченного прогресса. Там, где компьютерные программы имеют по большей части нисходящую организацию (прямые расчеты, шахматные программы, научные вычисления), компьютеры способны достичь огромной мощности и эффективности. В этих случаях компьютер также не нуждается в подлинном понимании выполняемых им действий, только здесь все необходимое понимание предварительно вложено в программу человеком (см. § 1.21).

Следует упомянуть и о том, что в системах нисходящего типа очень часты компьютерные ошибки, возникающие из-за ошибок в программах. Впрочем, такая ситуация является результатом человеческой ошибки, а это совершенно иное дело. Существуют — и порой даже приносят реальную пользу — автоматические системы исправления ошибок, однако они способны выловить далеко не все ошибки, некоторые оказываются им не по зубам.

Опасность чрезмерно доверчивого отношения к системам с полным компьютерным управлением хорошо иллюстрируется ситуациями, в которых упомянутая система в течение долгого времени работает вполне приемлемо, создавая, возможно, у человека *впечатление*, что она понимает, что делает. И вдруг неожиданно она выкидывает нечто совершенно безумное, что недвусмысленно показывает, что никакого подлинного понимания в ее действиях никогда не было (как в случае с неспособностью компьютера «Deep Thought» решить шахматную задачу, изображенную на рис. 1.7). Так что никогда не теряйте бдительности. Вооруженные знанием того, что «понимание» простонапросто не является вычислительным качеством, мы всегда должны помнить: никакой робот с компьютерным управлением таким качеством ни в коей мере обладать не может.

Разумеется, в отношении обладания способностью к пониманию люди и сами очень друг от друга отличаются. Как и компьютер, человек тоже может создать у окружающих впечатление присутствия в его действиях понимания, когда на самом деле никакого понимания там нет. Как правило, имеет место своего рода компромисс между подлинным пониманием, с одной стороны, и памятью и способностью к счету — с другой. Компьютеры сильны в последнем, но не способны достичь первого. Как хорошо известно преподавателям на всех уровнях (но, увы, *не всегда* известно правительственным чиновникам), гораздо более ценной во всех отношениях является способность к пониманию. Именно понимания (а не просто попугайского зазубривания правил и фактов) стремится добиться от своих учеников учитель. Одно из требований к составителю экзаменационных билетов (особенно в математике) как раз в том и заключается, чтобы по ответам абитуриента на вопросы можно было бы судить о его способности именно к пониманию, отдельно от способностей к запоминанию или счету — хотя эти последние, надо признать, также не лишены некоторой полезности.



## 8.3. Эстетика и т. д.

В вышеприведенных рассуждениях я говорил, по большей части, о способности к «пониманию», полагая ее существенным компонентом, напрочь отсутствующим в любой чисто вычислительной системе. Именно эта способность фигурировала в гёделевском рассуждении в § 2.5 — и именно ее отсутствие в бездумности вычислительного процесса продемонстрировало существенную ограниченность вычислений, побудив нас тем самым к поискам лучшего. И все же «понимание» — это лишь одна из способностей, за которые мы ценим сознательное восприятие. В более общем смысле мы, обладающие сознанием существа, получаем преимущество в любых обстоятельствах, где мы можем непосредственно «чувствовать» то, что нас окружает; и *этому* вычислительные системы не «научатся» никогда.

Меня спросят: каких же таких преимуществ оказывается *лишен* робот с компьютерным управлением в результате своей неспособности чувствовать? Что с того, что он не в состоянии оценить, скажем, ни красоту звездного неба, ни величественное великолепие Тадж-Махала тихим вечером, ни волшебных переплетений фуги Баха, ни даже суровой красоты теоремы Пифагора? Можно просто сказать, что робот много теряет, не будучи способным ощутить то, что ощущаем мы, сталкиваясь с такими проявлениями совершенства. Однако это далеко не весь ответ. Попробуем спросить иначе. Пусть робот действительно не способен ничего *чувствовать*, но нельзя ли запрограммировать компьютер таким хитроумным образом, чтобы он, тем не менее, смог создать великое произведение искусства?

Этот вопрос представляется мне чрезвычайно деликатным. Кратким ответом на него, думаю, будет «нет» — хотя бы по той

причине, что компьютер не способен испытывать чувственные ощущения, необходимые для того, чтобы отличить хорошее от плохого или превосходное от посредственного. Но тут можно задать встречный вопрос: а почему для того, чтобы вырабатывать собственные «эстетические критерии» и формировать собственные суждения, компьютер *непременно* должен обладать способностью «чувствовать»? Почему такие суждения не могут просто «возникнуть» после достаточно длительного обучения (восходящего типа)? Я, впрочем, думаю, что, как и в случае со способностью к пониманию, гораздо более вероятно, что упомянутые критерии все же придется в компьютер предварительно ввести, причем для получения этих самых критериев потребуется детальный нисходящий анализ, выполненный людьми (вполне возможно, не без помощи компьютера), в полной мере обладающими эстетическим чувством. Разработкой подобного рода схем занимались многие исследователи проблемы ИИ. Например, Кристофер Лонгет-Хиггинс (университет Суссекса) разработал несколько различных компьютерных систем, сочиняющих музыку согласно заложенным в них критериям. Еще в восемнадцатом веке Моцарт с современниками показали, как можно сочинять музыку с помощью так называемой «музыкальной игры в кости» — сочетая известные эстетически приятные фрагменты со случайными элементами, можно получать вполне сносные композиции. Аналогичные устройства были созданы и в области графических искусств — например, программа «AARON», разработанная Гарольдом Коэном, способна выдавать на гора в больших количествах «оригинальные» графические работы, генерируя случайные элементы и комбинируя их с имеющимися в ее распоряжении фиксированными образами в соответствии с определенными правилами. (Множество примеров такого «компьютерного творчества» можно найти в книге Маргарет Боден «Творческий разум» [32]; см. также [261].)

Думаю, что выражу общее мнение, если скажу, что среди продуктов такого рода деятельности пока нет ничего такого, что могло бы выдержать сравнение с любым творением умеренно способного художника-человека. Наверное, здесь уместно сказать, что даже при весьма значительных объемах предварительно введенных данных создаваемые компьютером «шедевры» оказываются напрочь лишены «души»! Иначе говоря, картина ничего не *выражает*, потому что компьютер ничего не *чувствует*.



Разумеется, случайно сгенерированная компьютерная работа может, просто по чистой случайности, оказаться и подлинным шедевром огромной художественной ценности. (Равно как и набирая буквы случайным образом, можно когда-нибудь получить «Гамлета».) В самом деле, следует признать, что и Природа способна волею случая сотворить настоящие произведения искусства, например, скалы причудливых очертаний или звезды в небе. Однако без способности *чувствовать* эту красоту невозможно отличить прекрасное от безобразного. Фундаментальная ограниченность полностью вычислительной системы проявится в полной мере еще в процессе *отбора*.

Опять же можно представить, что человек снабдит компьютер вычислительными критериями для такого различения, и это, возможно, какое-то время будет работать, коль скоро машине останется только генерировать очередные вариации на тему все того же эталона (возможно, так и создается большая часть рядовых «произведений» популярного искусства) — до тех пор, пока плоды такой деятельности не станут вызывать зевоту и нам не захочется чего-нибудь нового. На этом этапе машине потребуется какое-либо *подлинное* эстетическое суждение извне, чтобы выяснить, какие «новые идеи» имеют художественную ценность, а какие — нет.

Итак, помимо способности к *пониманию*, существуют и другие качества, каким полностью вычислительная система никогда не «научится» — например, способность к *эстетическому* восприятию. Сюда же, как мне представляется, следует отнести и все прочие качества и способности, что требуют осознания, — например, способность к *нравственному* суждению. Как мы убедились в первой части книги, суждение об *истинности* или *ложности* утверждения невозможно свести к чистому вычислению. То же применимо (возможно, даже с большей очевидностью) и к суждениям о *прекрасном* или о *добром*. Все эти способности требуют осознания и, как следствие, недоступны роботам с полностью компьютерным управлением. Для имитации роботом наличия этих способностей необходимо постоянное дополнительное управляющее воздействие со стороны какой-либо внешней, чувствующей и осознающей себя сущности — предположительно, человека.

Безотносительно к невычислительной природе упомянутых качеств, можно поинтересоваться, являются ли «красота» и «до-

брота» идеями *абсолютными* в платоновском смысле этого слова, где определение «абсолютный» применимо только к истине — в особенности, к математической истине. Сам Платон высказывался в поддержку такой точки зрения. Может быть, осознавая, мы каким-то образом связываемся с этими абсолютами, и именно в этом заключается уникальное предназначение сознания? Может быть, здесь и следует искать ключ к тому, *чем* наше сознание является в действительности и *для чего* оно нам дано? Не играет ли сознание роль своего рода «моста» между физическим миром и миром платоновских абсолютов? Эти вопросы мы еще раз затронем в последнем параграфе книги.

Вопрос об абсолютной природе нравственности имеет самое прямое отношение к юридическим проблемам, описанным в § 1.11. Некоторым образом связан с ним и вопрос о сущности «свободы воли», поставленный в конце § 1.11: возможно ли, что есть нечто, что не определяется наследственностью, влиянием окружения и всевозможными случайными факторами, — некая отдельная «самость», играющая ведущую роль в управлении нашими действиями? Я думаю, что мы пока еще очень далеки от ответа на этот вопрос. С полной уверенностью я могу утверждать (и аргументированно доказывать) лишь одно: что бы ни управляло в конечном счете нашим поведением, это что-то в принципе находится за пределами возможностей тех устройств, которые мы сегодня называем «компьютеры».

## 8.4. Опасности компьютерных технологий

Любые широкоприменяемые технологии несут с собой как блага, так и опасности. Так, помимо тех очевидных преимуществ, которые дают нам компьютеры, с быстрым развитием этой технологии связано и множество потенциальных угроз обществу. Одной из главных проблем, по-видимому, является чрезвычайная сложность всех совокупностей взаимосвязей, с которыми мы сталкиваемся благодаря компьютерам, — она приводит к тому, что ни один отдельно взятый индивидуум сегодня просто не в состоянии охватить разумом ни происходящее в целом, ни его последствия. И дело не только в самих компьютерах и их технических возможностях, но еще и в почти мгновенной глобальной связи между объединенными в сеть компьютерами по всему миру. Часть возможных проблем находит отражение в нестабильном поведении фондового рынка, где сделки теперь совершаются практически мгновенно на основании общемировых компьютерных прогнозов. Здесь, пожалуй, проблема заключается не столько в недостатке понимания каждым отдельным человеком всей взаимосвязанной системы как единого целого, сколько в нестабильности (не говоря уже о несправедливости), изначально заложенной в систему, которая идеально приспособлена для того, чтобы отдельные ее пользователи мгновенно сколачивали себе состояния путем опережения соперников в скорости счета или быстроте получения информации. Впрочем, вполне вероятно и то, что причиной различного рода нестабильностей и потенциальных опасностей станет одна лишь сложность системы как целого.



Подозреваю, что найдутся люди, которым возможный в недалеком будущем выход уровня сложности системы взаимосвязей за пределы человеческого понимания *не покажется* такой уж серьезной проблемой. Такие люди, возможно, верят в то, что когда-нибудь компьютеры и *сами* приобретут необходимое понимание системы. Однако, как мы могли убедиться, понимание отнюдь не относится к тем качествам, на которые компьютеры когда-либо окажутся *способны*, так что помощи с той стороны ждать не приходится.

Из одного лишь факта чрезвычайно быстрого развития компьютерных технологий (приводящего к тому, что компьютерная система чуть ли не на следующий день после своего появления на рынке становится морально устаревшей) вытекают и многие другие дополнительные проблемы. Необходимость в непрерывной модернизации и использование систем, зачастую не прошедших под давлением конкуренции надлежащих испытаний, — это лишь малая их часть, и в будущем ситуация вряд ли изменится к лучшему.

Глубинные же проблемы, с которыми мы только начинаем сталкиваться в новом высокотехнологичном, компьютеризованном и стремительно меняющемся мире, слишком многочисленны, и было бы безрассудством с моей стороны пытаться охватить их здесь все. Среди прочего в голову приходят разглашение частной информации, промышленный шпионаж и компьютерные диверсии. Еще одна тревожная возможность — «подделка» внешнего вида человека с целью использования, скажем, в телевизионной передаче для выражения мнений, какие «оригинал» ни в коем случае выражать не собирался[2]. Возникают и всевозможные социальные проблемы, не являющиеся непосредствен-

но компьютерными, но с компьютерами так или иначе связанные — например, благодаря способности компьютеров замечательно точно записывать и затем воспроизводить музыку и изображение, таланты небольшой избранной группы исполнителей можно без труда распространить по всему миру, что, вероятно, поставит в весьма невыгодное положение остальных, не столь именитых артистов. С аналогичной проблемой мы сталкиваемся и в случае с так называемыми «экспертными системами», позволяющими поместить мастерство и опыт нескольких избранных специалистов — скажем, от юриспруденции или медицины — в код компьютерной программы, что может привести к нанесению ущерба остальным практикующим врачам и юристам. Впрочем, думаю, что заменить специалиста-человека такие компьютерные экспертные системы вряд ли смогут (их удел — специалисту помогать), поскольку они не способны на *понимание*, которое может дать только личное общение.

Разумеется, есть у всех этих разработок и «светлая сторона» — если все сделано правильно. Плоды мастерства других (неважно, художников или ремесленников) сегодня более доступны, и их может оценить гораздо большее количество людей. Что касается проблемы сохранности частной информации, то уже сейчас существуют так называемые «шифры с открытым ключом» (см. [138]), которыми могут пользоваться как отдельные индивидуумы, так и небольшие компании (при этом не менее эффективно, нежели компании крупные), и которые, *по-видимому*, обеспечивают абсолютную защиту от «подслушивания». Использование таких шифров стало возможным лишь теперь, при наличии быстрых и мощных компьютеров — хотя эффективность этого способа шифрования до сих пор ограничена вычислительной сложностью факторизации больших чисел (возможно, здесь на смену обычным придут квантовые вычисления; некоторые идеи, указывающие на возможность создания в будущем квантовых компьютеров, изложены в §7.3, см. также [277, 278]). Как я упоминал в §8.1, возможно, что скоро для защиты от подслушивания мы будем использовать квантовую криптографию, эффективность которой также зависит от скорости выполнения значительных объемов вычислений. Очевидно, что нет однозначного способа оценить преимущества и опасности, порождаемые любой новой технологией, будь она непосредственно связана с компьютерами или нет.



В качестве заключительного комментария к таким компьютерно-социальным проблемам я хочу представить читателю небольшую вымышленную историю, которая в некотором роде выражает то беспокойство, которое я ощущаю в связи с возникновением целой новой области потенциальных проблем. Насколько мне известно, об этом новом классе «компьютерных» опасностей еще никто не говорил, однако мне они представляются весьма серьезными.

## 8.5. Неправильные выборы

Приближается день долгожданных выборов. На протяжении последних недель были проведены многочисленные опросы общественного мнения. Результаты почти единодушно предсказывают отставание правящей партии по голосам на три-четыре процента. Как и ожидалось, имеются некоторые колебания и отклонения от этой цифры в ту или иную сторону — ожидалось, поскольку цифры в опросах базируются на относительно малых выборках (где-то в пределах нескольких сотен избирателей за раз), тогда как по населению в целом (несколько десятков миллионов человек) наблюдаются существенные изменения от места к месту. В самом деле, предел погрешности каждого из опросов и сам может составить те самые три-четыре процента, так что ни на один из опросов в действительности полностью положиться нельзя. И все же в совокупности свидетельства производят куда более выгодное впечатление. Взятые вместе, результаты опросов демонстрируют гораздо меньшую погрешность, а согласие между ними нарушается как раз таким разбросом, какой предсказывает статистическая теория. Усредненным результатам теперь, наверное, вполне можно доверять, причем погрешность составляет менее двух процентов. Поговаривают, правда, что в канун дня выборов цифры в опросах заметно сместились в пользу правящей партии; а в сам день выборов кое-кого из ранее воздержавшихся (или даже из активных противников) вполне могут «уговорить» отдать-таки правящей партии свой голос. Однако даже если так, это смещение не принесет правящей партии большой пользы, разве что полученный в результате отрыв от ближайшего соперника составит не менее 8% голосов, поскольку только в этом случае правящая партия получит то минимальное большинство

голосов, которое необходимо для того, чтобы предотвратить объединение своих противников в коалицию. Впрочем, опросы — это всего лишь предварительные прикидки, разве нет? Только *подлинное* голосование выразит действительную волю народа, а какова эта воля, мы узнаем из подсчета голосов в день выборов.

День выборов настал... и прошел. Голоса подсчитали, и результат почти для всех оказался полной неожиданностью — особенно для организаций, проводивших опросы и вложивших в них так много сил и умения (не говоря уже о репутации). Правящая партия остается у власти, получив вполне удовлетворительное большинство голосов — те самые 8% преимущества над ближайшими соперниками. Огромное количество избирателей пребывает в полном недоумении — и даже в ужасе. Другие, хотя и удивлены не меньше, но весьма довольны. Однако результаты выборов не соответствуют истине. Они были фальсифицированы с помощью хитроумных средств, и никто ничего не заметил. Заранее наполненных урн для голосования там не было, бюллетени никто не терял, не подменивал и не дублировал. Люди, занятые в подсчете голосов, сделали свое дело добросовестно и по большей части без ошибок. И все же результаты выборов оказались чудовищно подтасованы. Как же так получилось, и кто это сделал?

Не исключено, что весь кабинет правящей партии в полном составе понятия не имеет о том, что произошло. Не факт, что кто-то из них является непосредственным виновником преступления, однако в выигрыше в результате оказываются они все. За кулисами скрываются другие, те, кто имеет основания опасаться за собственное существование, если правящей партии случится потерпеть поражение. Они входят в состав некоей организации, которая пользуется бóльшим доверием у правящей партии (и не без причины!), чем у оппозиции, — партия не только строго и бережно хранит тайну темных делишек этой организации, но и способствует расширению ее деятельности. Хотя сама организация вполне законна, многое из того, чем она занимается, законным не назовешь, не чурается она и незаконных политических игр. Возможно, члены организации искренне (заблуждаться тоже можно искренне) опасаются, что противники правящей партии разрушат страну или даже «предадут» ее во имя чуждых идеалов иностранных держав. Есть в организации и свои эксперты — непревзойденные мастера — в области создания компьютерных вирусов!



Помните, что способен натворить компьютерный вирус? Ближе всего нам знакомы те, что в некий заранее назначенный день уничтожают всю информацию на дисках компьютера, этим вирусом зараженного. Бывает так, что пользователь сидит и с ужасом наблюдает, как буквы на дисплее его компьютера ссыпаются со своих мест в нижнюю часть экрана и исчезают. Бывает, на экране появляется какое-нибудь непристойное сообщение. В любом случае данные могут оказаться потерянными безвозвратно. Более того, если вставить в такой компьютер дискету и попробовать ее прочитать, то дискета тоже подхватит заразу и передаст ее при случае на другой компьютер. Замеченный вирус можно, в принципе, уничтожить с помощью антивирусной программы, но только в том случае, если природа вируса известна заранее. Если же вирус успел нанести удар, то поделать уже ничего нельзя.

Такие вирусы обычно создают хакеры-любители, зачастую этими хакерами становятся разочаровавшиеся в жизни программисты, желающие кому-нибудь насолить по тем или иным причинам, иногда вполне объяснимым, иногда нет. Однако члены упомянутой организации — отнюдь не любители; им немало платят, и в своей области они настоящие профессионалы. Возможно, многие из их действий продиктованы подлинной заботой об интересах родной страны, но бывает, несомненно, и так, что по указанию своих непосредственных начальников они делают вещи, менее простительные с точки зрения морали. Созданный программистами организации для известной цели вирус невозможно засечь стандартными антивирусными программами, и сработать он должен лишь однажды, в заранее назначенный день — вождь правящей партии, конечно же, знает, на какой день назначены выборы, знают об этом и те, кому вождь доверяет. После того, как задание будет выполнено, — а на этот раз задание предстоит куда более тонкое, чем просто стереть данные, — вирус самоуничтожится, не оставив после себя ни единого следа, если не считать, разумеется, самого преступления.

Для того, чтобы такой вирус надлежащим образом сработал на выборах, необходимо, чтобы какой-то этап в подсчете голосов происходил без участия людей (считающих либо вовсе без применения техники, либо с помощью карманного калькулятора). (Вирус может инфицировать только универсально программируемые компьютеры.) Допустим, содержимое отдельных

урн считают люди и считают правильно; однако результаты этих подсчетов необходимо складывать. Насколько же эффективнее, точнее, да и современнее складывать эти числа — а их там, может быть, сотни — на компьютере, нежели вручную или с помощью калькулятора! Разумеется, никаких ошибок здесь просто быть не может. Чей бы компьютер ни использовался для подсчета общей суммы, результат будет одинаковым. Члены правящей партии получат в точности тот же результат, что и их главные противники, равно как и любая из третьих заинтересованных партий или вовсе нейтральный наблюдатель. Они даже могут использовать компьютеры разных моделей или марок, на результат это никоим образом не повлияет. Экспертам нашей зловещей организации об этих разных компьютерах известно все — и для каждого заготовлен свой вирус. По своей структуре вирусы для разных систем несколько отличаются друг от друга, однако последствия их «работы» будут одинаковыми, а согласие между результатами, полученными с помощью различных машин, убедит даже самых упрямых скептиков.

Несмотря на то, что все машины дадут одинаковые цифры, цифры эти все до единой будут неверными. Все цифры хитроумно фабрикуются в соответствии с некоей точной формулой, зависящей до некоторой степени от реального распределения голосов, — отсюда, согласие между результатами, полученными с помощью различных компьютеров, и смутное правдоподобие этих самых результатов, — с тем, чтобы дать правящей партии именно то преимущество, в котором она нуждается; и хотя доверчивость избирателей при этом подвергается некоторому испытанию, общий результат представляется вполне приемлемым. Все *выглядит* так, будто значительное число избирателей в последнюю минуту решило проявить осторожность и проголосовать за правящую партию.

В гипотетической ситуации, описанной в этой истории, избиратели на самом деле вовсе не передумывали в последний момент, и результаты выборов оказались весьма далеки от истинного положения дел. Хотя на написание ее меня вдохновили наши последние (1992 год) выборы в британский парламент, я должен особо подчеркнуть, что официально принятая в Великобритании система подсчета голосов возможность такого рода мошенничества *полностью исключает*. На всех этапах подсчет выполняется вручную. Может, конечно, показаться, что этот метод неэффективен и давно устарел, однако отказываться от него еще, как мне представляется, рано — по крайней мере, до тех пор, пока не будет создана система, снабженная простыми и исключающими малейшее подозрение средствами защиты от подобного мошенничества.



С другой, более положительной, стороны, современные компьютеры предлагают замечательные возможности для введения систем голосования, в которых мнение избирателей будет представлено гораздо объективнее, чем сейчас. Здесь, разумеется, не место вдаваться в подробное обсуждение этих вопросов, однако суть такова, что новые системы позволяют избирателю не просто отдать свой голос за одного-единственного кандидата, но сообщить и множество иных сведений. Все эти сведения компьютерная система способна проанализировать мгновенно, и результат можно будет получить сразу же после окончания процедуры голосования. Однако, как показывает рассказанная выше история, применять такую систему следует крайне осторожно, даже если в ней предусмотрены всесторонние и общедоступные проверки, убедительно предотвращающие любое такое техническое мошенничество.

Осторожность следует проявлять не только на выборах; «вирусный» метод можно применить и в других ситуациях, например, подпортить банковские счета компании-соперника. Можно придумать множество различных способов вредоносного использования специально разработанных, незаметных и коварных компьютерных вирусов. Надеюсь, что моя история убедит читателей в том, что все действия компьютеров — даже самые очевидных действия даже самых надежных компьютеров — должны постоянно контролироваться человеком. И дело здесь не столько в том, что компьютеры ничего не *понимают*, сколько в том, что они крайне подвержены манипуляциям со стороны тех немногих людей, кто *понимает* все тонкости специфики их программирования.

## 8.6. Физический феномен сознания

Во второй части книги мы, не выходя за пределы научно объяснимого, попытались отыскать, если можно так выразиться,

место в физике, пригодное для размещения субъективного опыта. Как выяснилось, для успеха такого поиска сегодняшние границы научного понимания придется расширить. Я почти не сомневаюсь в том, что то фундаментальное изменение, которому неминуемо должна подвергнуться наша традиционная картина физической реальности, придет откуда-то со стороны феномена редукции квантового состояния. Прежде чем физика сможет смириться с чем-то, настолько чуждым всем современным физическим представлениям, как феномен сознания, следует ожидать полного пересмотра самих основ всех существующих философских воззрений на природу реальности. По этому поводу у меня есть кое-какие краткие замечания, которые я приведу очень скоро — в следующем, последнем, параграфе. А пока давайте попробуем ответить на несколько более простой вопрос: где в известном физическом мире, учитывая предложенные на этих страницах доказательства, можно надеяться отыскать сознание?

Необходимо с самого начала внести полную ясность: выводы из упомянутых доказательств и прочих моих рассуждений носят, по большей части, «отрицательный» характер. Мы убедились, например, что современные компьютеры сознанием *не обладают*, но мы по-прежнему слабо представляем себе, что именно в объекте приводит к возникновению у него сознания. Основываясь на собственном опыте, мы полагаем (по крайней мере, пока), что феномен этот обычно присущ биологическим структурам. На одном конце шкалы у нас люди, и тут, конечно же, сомнений почти нет — что бы ни представляло собой в действительности сознание, оно, в нормальном своем состоянии, так или иначе связано с бодрствующим (а возможно, и со спящим) человеческим мозгом.

Что же мы видим на другом конце шкалы? Я убежден, что фокус нашего внимания следует переместить с нейронов на микротрубочки цитоскелета: именно там, вероятнее всего, возникают коллективные (когерентные) квантовые эффекты — а без такой квантовой когерентности не будет и новой **OR**-физики, которая, как мне представляется, должна стать необходимым невычислительным условием для объяснения феномена сознания в научных терминах. Однако цитоскелеты есть у всех эукариотических клеток — клеток, из которых состоят растения и животные; эукариотами являются и одноклеточные организмы, такие как парамеции и амебы, но не бактерии. Следует ли из этого, что парамеция также обладает некоторым зачаточным сознанием?



Возможно ли, чтобы парамеция «знала» (в любом смысле этого слова), что делает? А что же *отдельные* клетки человеческого тела — клетки мозга, например, или клетки печени? Может быть, когда мы поймем физическую природу процесса осознания настолько хорошо, что будем в состоянии ответить на эти вопросы, нам придется признать, что ничего такого уж нелепого в этих предположениях нет. Я не знаю. *Знаю* я лишь то, что проблема эта является целиком и полностью *научной*, а это значит, что когда-нибудь решение неизбежно будет найдено, вне зависимости от того, насколько далеки мы от этого решения сейчас.

Иногда утверждают — исходя из общих философских принципов, — что узнать, обладает ли способностью к осознанию какое бы то ни было существо, отличное от тебя самого, принципиально невозможно, не говоря уже о том, чтобы выяснить, нет ли каких-нибудь зачатков сознания у парамеции. Думаю, такая позиция чересчур узка и пессимистична. В конце концов, когда речь идет об установлении факта наличия у некоего объекта того или иного физического свойства, никто же не настаивает на *абсолютной уверенности*. Настанет время, и на вопросы, касающиеся способности к осознанию, мы будем отвечать с той же степенью уверенности, с какой сегодняшние астрономы высказываются о небесных телах, удаленных от нас на многие световые годы. Еще совсем недавно ученые утверждали, что нам никогда не узнать, из чего состоят Солнце и звезды и что находится на обратной стороне Луны. Сегодня у нас есть подробные карты обратной стороны Луны (фотосъемка из космоса), а состав Солнца изучен до мельчайших подробностей (наблюдение линий солнечного спектра, а также тщательное и подробное моделирование физических процессов, происходящих внутри Солнца). Известен нам и подробный состав далеких звезд, причем с очень хорошей точностью. Мы можем даже сказать (и в некоторых отношениях — сказать точно), из чего состояла вся Вселенная на начальных этапах ее развития (см. конец § 4.5).

Однако в отсутствие необходимых теоретических идей суждения относительно обладания сознанием не выходят (по большей части) из разряда предположений. Мое собственное предположение по этому поводу таково: с некоторых пор я совершенно уверен, что на планете Земля сознание не является *исключительной* прерогативой человека. В одной из наиболее захватывающих телевизионных программ Дэвида Аттенборо[3] был эпи-

зод, после просмотра которого зрителям было трудно не поверить не только в то, что слоны, например, способны на сильные чувства, но и в то, что чувства эти не так уж далеки от тех, из каких в человеческих обществах возникают религии. Вожак стада — самка, потерявшая около пяти лет назад сестру, — ведет стадо на место ее гибели, значительно отклоняясь от обычного маршрута; прибыв на место и обнаружив останки, вожак очень осторожно поднимает с земли череп, а затем слоны начинают передавать его друг другу, поглаживая хоботами. То, что слоны способны и на понимание, убедительно, хотя и жутковато, показано в другой телевизионной программе[4]. Фильм, отснятый с вертолета, участвующего в операции, деликатно называемой «отбраковкой», очень хорошо передает ужас, охватывающий слонов, когда они до конца осознают, что происходит, и понимают, что никто из стада живым отсюда не уйдет.

Множество свидетельств имеется и в пользу наличия сознания (и самосознания) у человекообразных обезьян, и я почти не сомневаюсь, что феномен сознания присущ и животным формам, значительно менее «высокоорганизованным». Например, в еще одной телевизионной программе[5] — рассказывающей о чрезвычайной ловкости, решительности и изобретательности белок (некоторых) — меня особенно поразил фрагмент, в котором белка сообразила, что перекусив проволоку, она сможет освободить контейнер с орехами, подвешенный на некотором расстоянии от нее. Вряд ли этот акт понимания был инстинктивным или вытекал из какого-то прошлого опыта белки. Для того, чтобы оценить, насколько положительным окажется результат ее действия, белка должна была понять хотя бы на элементарном уровне *топологию* всей конструкции (см. также § 1.19). Мне представляется, что в данном случае мы наблюдали проявление подлинного *воображения* — а для этого, разумеется, необходимо сознание!

Почти не остается сомнений и в том, что сознание может «присутствовать» в разной степени — между «в полном сознании» и «без сознания» возможны и другие состояния. О себе, например, я могу сказать совершенно определенно: иногда я чувствую себя более «в сознании», иногда — менее (скажем, во время сна сознание присутствует в гораздо меньшей степени, чем когда я бодрствую).

Насколько же далеко мы должны зайти в наших поисках? На этот счет существуют самые различные мнения. Что касается



меня, то я с трудом представляю себе, что сознанием (в какой бы то ни было степени) могут обладать насекомые — особенно после того, как я увидел документальный фильм о жизни насекомых, где было показано, как некий жук с жадностью пожирает другого жука, совершенно, по всей видимости, не обращая внимания на то, что его самого в это время ест третий. Тем не менее, как упоминалось в § 1.15, поведение простого муравья отличается чрезвычайной сложностью и точностью. Надо ли полагать, что замечательно эффективные управляющие системы муравья работают вовсе без участия того принципа (каким бы он ни был), благодаря которому мы сами получаем способность понимать? Управляющие нейроны муравья также не лишены цитоскелетов, и если в этих цитоскелетах имеются микротрубочки, способные поддерживать квантовокогерентные состояния, которые, согласно моему предположению, играют ключевую роль в процессе осознания, то не следует ли из этого, что муравей является счастливым обладателем того же самого неуловимого сознания, что и мы с вами? Если же микротрубочки в человеческом мозге и в самом деле обладают той неимоверной сложностью, что необходима для поддержания коллективных квантовокогерентных процессов, то не совсем понятно, почему естественный отбор развил такую способность только в нас и в наших ближайших многоклеточных родственниках (в некоторых из них, по крайней мере). Такие квантовокогерентные состояния могли оказаться весьма полезными и для первых эукариотических одноклеточных, хотя в чем эта полезность могла бы состоять, мы можем только предполагать.

Одной лишь макроскопической квантовой когерентности для возникновения сознания, разумеется, *недостаточно* — иначе сознанием обладали бы и сверхпроводники! Однако вполне вероятно, что такая когерентность является *частью* того, что для сознания *необходимо*. Мозг обладает чрезвычайно сложной организацией, и поскольку сознание, по-видимому, представляет собой результат *глобальной* координации всевозможных мыслительных процессов, следует искать когерентность в масштабах, гораздо более крупных, нежели отдельные микротрубочки или даже целые цитоскелеты. Должна существовать существенная квантовая сцепленность между состояниями, поддерживаемыми внутри отдельных цитоскелетов во многих нейронах, — т. е. нечто вроде коллективного квантового состояния, охватывающего об-

ширные области мозга. Однако и этого недостаточно. Для того, чтобы в системе могли происходить какие бы то ни было полезные невычислимые процессы — что я считаю существенной частью сознания, — необходимо, чтобы система была способна специфическим образом задействовать подлинно *неслучайные* (невычислимые) аспекты **OR**-процедуры. Предположение, которое я сделал в § 6.12, дает нам (по крайней мере) некоторое представление о соответствующих *масштабах*, начиная с которых можно говорить о каком-то существенном действии точной и математически невычислимой **OR**-процедуры.

Таким образом, предложенные мною в настоящей книге соображения дают в некотором роде основу для высказывания правдоподобных *догадок* (пока, во всяком случае) относительно уровня, на котором можно ожидать возникновения способности к осознанию. Процессы, которые могут быть адекватно описаны в рамках вычислимой (или случайной) физики, не могут, согласно моей точке зрения, иметь отношения к сознанию. С другой стороны, даже существенное участие точной невычислимой **OR**-процедуры само по себе вовсе не обязательно *подразумевает* наличие сознания — хотя и является, на мой взгляд, *необходимым* для этого условием. Разумеется, критерию не достает определенности, однако ничего лучшего на данный момент у меня нет. Посмотрим, далеко ли он нас заведет.

Будем исходить из сделанных в § 6.12 предположений относительно того, где должна проходить граница между классическим и квантовым уровнями, а также из изложенных в §§ 7.5–7.7 биологических умопостроений, согласно которым эту границу, возможно, следует искать где-то в области сопряжения внутренних и внешних процессов в системах микротрубочек клетки или совокупности клеток. В качестве существенного дополнения заметим, что если редукция вектора состояния происходит просто потому, что рассматриваемая система оказывается сцеплена с слишком большим объемом окружения, то процедуру **OR** можно считать эффективно *случайным* процессом, для описания которого вполне пригодна стандартная FAPP-аргументация (представленная в общих чертах в § 6.6); процедура **OR** в данном случае полностью идентична процедуре **R**. Необходимо, чтобы эта редукция происходила в точности тогда, когда начинают действовать невычислительные (и пока неизвестные) *правила* нашей гипотетической **OR**-теории. Хотя об этих правилах мы ничего не знаем, мы можем (по крайней мере, в принципе) составить некоторое представление о том уровне, на котором теория начинает соответствовать реальности. Таким образом, для того, чтобы упомянутые невычислимые аспекты процедуры **OR** смогли сыграть свою роль, необходимо, чтобы та или иная квантовая когерентность поддерживалась до тех пор, пока перемещение вещества (вследствие взаимодействия между внутренними и внешними микротрубочковыми процессами) не достигнет определенного предела, *как раз* достаточного для того, чтобы **OR**-процедура произошла *прежде*, чем успеет вмешаться случайное окружение.



Что касается микротрубочек, то я предлагаю следующую картину: *внутри* трубок происходят «квантовокогерентные колебания», слабо связанные с вычислительной «клеточноавтоматной» активностью, обусловленной конформационными переходами димеров тубулина *на внешней* поверхности трубок. Пока квантовые колебания остаются изолированными, уровень для **OR** слишком низок. Однако, поскольку процессы внутри и снаружи связаны, квантовое состояние вскоре захватывает тубулины, и на некотором этапе происходит редукция (**OR**). Необходимо, чтобы **OR** происходила *прежде*, чем с квантовым состоянием окажется сцеплено микротрубочковое окружение, потому что как только возникает такая сцепленность, невычислимые аспекты **OR**-процедуры теряются, и она превращается в «обычную» **R**-процедуру.

Итак, остается лишь выяснить, достаточна ли конформационная активность тубулина в отдельной клетке (в парамеции, например, или в клетке человеческой печени) для того, чтобы обусловленное ею перемещение масс удовлетворило бы критерию из § 6.12 и процедура **OR** произошла бы именно тогда, когда нужно, или же этой активности недостаточно, и **OR** задержится до тех пор, пока окружение и в самом деле не возмутится, — и игра (призом в которой невычислимость) будет проиграна. Судя по первому впечатлению, так оно и есть — конформационная активность тубулина перемещает слишком малое количество вещества, и на требуемом уровне никакой **OR**-процедуры не происходит. Если же клеток много, ситуация выглядит гораздо более многообещающей.

Возможно, глядя на такую картину (в ее теперешнем виде) действительно не остается ничего другого, как предположить, что невычислительные условия для появления сознания могут воз-

никнуть только в больших совокупностях клеток, что мы и имеем в случае достаточно большого мозга(6). Однако я порекомендовал бы соблюдать (по крайней мере, на данном этапе) известную осторожность. Как физические, так и биологические аспекты предлагаемой картины сформулированы слишком приблизительно, чтобы можно было прямо сейчас делать какие-то однозначные выводы в отношении следствий из той точки зрения, которую я здесь представляю. Очевидно, что даже с учетом рассмотренных выше конкретных предложений потребуется еще немало исследований, как физических, так и биологических, прежде чем мы сможем сделать сколько-нибудь обоснованное предположение относительно места сознания в материальном мире.

Следует обратить внимание и на некоторые другие вопросы. Например, какая часть мозга действительно задействована в поддержании состояния сознания? Вероятнее всего, *весь* мозг для этого не требуется. Похоже на то, что многие функции мозга с сознанием никак не связаны. Взять хотя бы мозжечок (см. § 1.14), который, как это ни поразительно, работает абсолютно *бессознательно*. Именно мозжечок отвечает за координацию и точность наших действий в тех случаях, когда эти самые действия выполняются без участия сознания (см., например, НРК, с. 379–381). Из-за полной бессознательности его функций мозжечок часто называют «просто компьютером». Было бы, несомненно, весьма поучительно выяснить, есть ли какие-нибудь различия (и если есть, то какие именно) между клеточной или цитоскелетной организациями мозжечка и головного мозга, поскольку именно с последним, по всей видимости, гораздо более тесно связано сознание. Интересно, что если судить лишь по количеству нейронов, то разница между мозгом и мозжечком невелика — в мозге нейронов всего лишь в два раза больше, чем в мозжечке, причем отдельные клетки в мозжечке образуют, в общем случае, значительно больше синаптических связей, чем клетки мозга (см. § 1.14 и рис. 1.6). Очевидно, простым подсчетом нейронов тут не обойтись, следует искать глубже[1].

[1] Поскольку в нейроанатомии я человек вполне посторонний, меня не мог не поразить факт наличия в организации мозга одной особенности (похоже, так и не нашедшей до сих пор объяснения), которой мозжечок не обладает. Бóльшая часть сенсорных и двигательных нервов «идут наперекрест», т. е. левая сторона мозга отвечает в основном за правую сторону тела, и наоборот. И не только это — та область мозга, что обрабатывает зрительные образы, находится сзади, а та, что заведует ногами, находится вверху; так же обстоит дело и с ушами: сигналы из правого уха обрабатываются слева, а из левого — справа. Нельзя сказать, что эта особенность мозга носит абсолютно универсальный характер, но я не могу отделаться от ощущения, что это не случайно. Потому что мозжечок устроен иначе. Может ли быть так, что сознание каким-то образом выигрывает от того, что нервным сигналам приходится идти «длинной дорогой»?



Возможно, что-либо поучительное удастся извлечь и из изучения процесса «научения», посредством которого движения, первоначально осознаваемые мозгом, переходят под бессознательный мозжечковый контроль. Не исключено, что «обучающие процедуры» мозжечка окажутся очень похожими на те, с помощью которых приверженцы коннекционистской философии обучают искусственные нейронные сети. Впрочем, даже если так оно и есть и даже если верно *также* то, что в терминах таких процедур можно объяснить (хотя бы частично) работу *мозжечка* — что подразумевается, например, в коннекционистском подходе к исследованию зрительной коры(7), — нет никаких оснований полагать, что то же непременно окажется верно и в случае тех аспектов деятельности головного мозга, которые связаны с сознанием. В самом деле, как свидетельствуют представленные в первой части книги доказательства, для объяснения высших когнитивных функций, непосредственно связанных с сознанием, необходимо нечто, в корне отличное от коннекционизма.

## 8.7. Три мира и три загадки

Попробуем свести все вышесказанное вместе. На протяжении всей книги мы пытаемся найти ответ на главный вопрос: как можно соотнести феномен сознания с нашим научным мировоззрением? Надо признать, я мало что могу сказать о сознании вообще. Поэтому я сосредоточился (в первой части) на одном частном ментальном качестве: способности к *сознательному пониманию*, в частности, к математическому пониманию. Только на примере этого ментального качества я смог достаточно убедительно показать, что возникновение способности к пониманию в результате какой бы то ни было чисто вычислительной активности решительно *невозможно*, вычислением нельзя даже адекватно моделировать такую способность — особо следует отметить, что ничто в моих рассуждениях не указывает и на то, что *математическое* понимание в чем бы то ни было принципиаль-

но отличается от прочих видов понимания. Отсюда вывод: какая бы активность мозга ни отвечала за сознание (по крайней мере, в этом конкретном его проявлении), она должна основываться на физических процессах, описать которые численное моделирование неспособно. Во второй части мы попытались найти область в науке для соответствующего физического процесса, действительно способного вывести нас за пределы чистой вычислительности. Для того чтобы охватить встающие перед нами при этом фундаментальные проблемы, я воспользуюсь в дальнейшем метафорой трех различных миров и трех «великих загадок», связывающих эти миры вместе. Миры в чем-то похожи на те, что описывал Поппер (см. [309]), однако акценты я расставляю совершенно иначе.

Наиболее близок нам *мир наших сознательных восприятий* — знание об этом мире мы получаем самым непосредственным образом и о нем же мы знаем меньше всего в смысле точного научного описания. В этом мире есть счастье, боль и цвет. В нем хранятся наши самые ранние детские воспоминания и ждет своего часа страх смерти. В нем — любовь, понимание, знание различных фактов, а также невежество и мстительность. Этот мир содержит образы столов и стульев, здесь запахи, звуки и всевозможные ощущения смешиваются с нашими мыслями и решимостью действовать.

Известны нам и два других мира — не так непосредственно, как мир восприятий, но зато об этих мирах мы знаем довольно много всего. Один из них мы называем *физическим миром*. В нем находятся настоящие столы и стулья, телевизоры и автомобили, люди, человеческие мозги и импульсы нейронов. В этом мире есть Солнце, Луна и звезды. В нем же — облака, ураганы, скалы, цветы и бабочки, а на более глубоком уровне — молекулы и атомы, электроны и фотоны, время и пространство. Еще там есть цитоскелеты, димеры тубулина и сверхпроводники. Не совсем ясно, почему мир восприятий должен иметь что-то общее с физическим миром, однако, судя по всему, так оно и есть.

Что касается второго мира из упомянутых двух, то само его существование многими ставится под сомнение. Речь идет о *платоновском мире математических форм*. Здесь обитают натуральные числа 0, 1, 2, 3, ... и алгебра комплексных чисел. Здесь мы найдем теорему Лагранжа о том, что любое натуральное число есть сумма четырех квадратов, и самую знаменитую из



теорем евклидовой геометрии — теорему Пифагора (о квадратах сторон прямоугольного треугольника). Где-то здесь находится правило $a \times b = b \times a$ для любых натуральных чисел и тот факт, что означенное правило не работает в случае «чисел» некоторых других типов (например, тех, что участвуют в грассмановом произведении, упомянутом в § 5.15). Этот же платоновский мир содержит геометрии, отличные от евклидовой, геометрии, в которых теорема Пифагора неверна. Здесь есть бесконечность и невычислимость, рекурсивные и нерекурсивные ординалы. Здесь — незавершаемое действие машины Тьюринга и машина с оракулом, а также многие классы математических задач, неразрешимые вычислительными методами, такие как задача о замощении плоскости плитками полиомино. В этом мире мы встретим электромагнитные уравнения Максвелла и гравитационные — Эйнштейна, равно как и бесчисленные удовлетворяющие им теоретические пространства-времена, как реалистичные физически, так и совершенно невероятные. Именно здесь пребывают математические модели столов и стульев, которыми можно воспользоваться в «виртуальной реальности», а также модели черных дыр и ураганов.

Имеем ли мы право утверждать, что платоновский мир действительно является «миром» — миром, который «существует» в том же смысле, в каком существуют прочие два мира? Читателю, возможно, покажется, что это вовсе не мир, а просто какой-то пыльный склад для абстрактных концепций, которые понапридумывали математики. Однако существование мира математических идей опирается на фундаментальный, вневременной и универсальный характер этих самых идей и на тот факт, что описываемые ими законы никоим образом не зависят от тех, кто их открыл. Этот «склад» (если это и впрямь склад) построен не нами. Натуральные числа были в этом мире задолго до того, как на Земле появились первые человеческие существа — да и все остальные существа, если уж на то пошло, — и останутся после того, как вся жизнь во Вселенной исчезнет. То, что любое натуральное число есть сумма четырех квадратов, было истиной всегда, а вовсе не стало ею вдруг после того, как Лагранж призвал из небытия соответствующую теорему. Натуральные числа, настолько большие, что оказываются не по зубам любому компьютеру, какой вы можете вообразить, все равно являются суммами четырех квадратов, пусть даже мы никогда и не узнаем, квад-

ратов каких именно чисел. Всегда будет истинным утверждение, что общей вычислительной процедуры для установления факта незавершаемости действия машины Тьюринга не существует, и оно всегда было истинным, задолго до того, как Тьюрингу пришло в голову его определение вычислимости.

Тем не менее, многие возражают, утверждая, что абсолютный характер математической истины никоим образом не является аргументом в пользу реальности «существования» математических концепций и математических истин. (Время от времени я слышу, что математический платонизм якобы устарел. Разумеется, мне известно, что сам Платон умер что-то около 2340 лет назад, однако едва ли это можно считать достаточной причиной! Более серьезную причину могут составить трудности, с которыми порой сталкиваются философы, пытаясь обосновать целиком и полностью абстрактный мир, способный оказывать реальное воздействие на мир физический. Эта фундаментальная проблема, собственно, является частью одной из тех загадок, к которым мы очень скоро перейдем.) На деле же идея реальности математических концепций вполне естественна для математиков, чего нельзя сказать о тех, кто никогда не испытывал радости исследования чудес и тайн того мира. Впрочем, на данном этапе от читателя не требуется соглашаться с тем, что математические концепции действительно образуют «мир», реальность которого сравнима с реальностью физического и ментального миров. Различия во взглядах на природу математических концепций для нас пока существенной роли не играют. Можете, если хотите, рассматривать «платоновский мир математических форм» как риторическую фигуру, введенную для удобства последующих рассуждений. Когда мы доберемся до трех загадок, связывающих эти три «мира», причина именно такого выбора слов, возможно, станет несколько яснее.

Что же это за загадки? Для начала взгляните на рис. 8.1. Первая загадка: почему столь точные и фундаментальные математические законы играют такую важную роль в поведении физического мира? Кажется, что сам мир физической реальности каким-то таинственным образом возникает из платоновского мира математики. Этот процесс проиллюстрирован направленной вниз стрелкой на рисунке справа — от платоновского мира к физическому. Вторая загадка: как физический мир порождает восприятие объектов в сознании? Каким таким таинственным образом сложно организованные материальные объекты производят из самих себя объекты ментальные? Этот процесс представлен на рис. 8.1 стрелкой внизу, направленной от физического к ментальному миру. И наконец, последняя загадка: как мысль «творит» из той или иной ментальной модели математическую концепцию? Эти по виду нечеткие, ненадежные и часто вовсе неподходящие ментальные инструменты, доставшиеся нам, похоже, в комплекте с ментальным миром, каким-то таинственным образом оказываются, тем не менее, способны (по крайней мере, когда они «в ударе») производить из пустоты абстрактные математические формы, открывая нам тем самым доступ, через посредство понимания, в платоновское царство чистой математики. Этот процесс символизирует стрелка слева на рисунке, направленная вверх, от ментального миру к платоновскому.



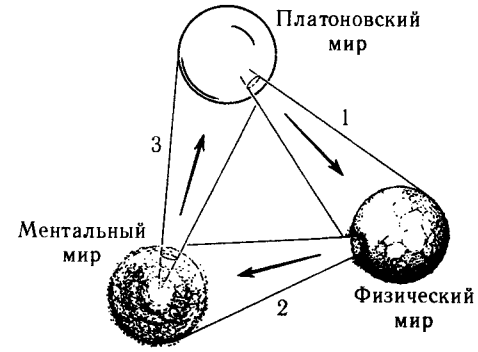



Рис. 8.1. Кажется, что каждый из трех миров — платоновский математический, физический и ментальный — неким таинственным образом «произрастает» из какой-то малой части своего предшественника (или, по крайней мере, очень тесно с этим предшественником связан).

Сам Платон большое внимание уделял первой из этих стрелок (а также, на свой лад, третьей), и неустанно подчеркивал различие между совершенной математической формой и ее несовершенной «тенью» в физическом мире. Так, сумма углов математического треугольника (евклидова треугольника, обязательно уточним мы сегодня) составляет ровно два прямых угла, тогда

как углы физического треугольника, сделанного, скажем, из дерева со всей точностью, на которую мы способны, образуют в сумме угол, величина которого очень близка к требуемой, но все же не равна ей. Эти свои соображения Платон изложил в виде притчи. Он вообразил нескольких граждан, заточенных в пещере и прикованных таким образом, чтобы они не могли видеть находившихся за их спинами совершенных форм, отбрасывающих в свете костра тени на стену пещеры, доступную взорам прикованных граждан. Таким образом, люди непосредственно видели лишь несовершенные тени тех форм, к тому же искаженные неровным светом костра. Совершенные формы символизировали собой математические идеи, а тени на стене — мир «физической реальности».

Со времен Платона основополагающая роль математики в объяснении воспринимаемой структуры и действительного поведения физического мира возросла чрезвычайно. В 1960 году видный физик Юджин Вигнер прочел знаменитую лекцию под названием «Непостижимая эффективность математики в физических науках». В ней он отметил поразительную точность и хитроумную применимость замысловатых математических конструкций, которые физики регулярно и во все бóльших количествах обнаруживают в своих описаниях реальности.

Для меня наиболее впечатляющим примером эффективности математики является общая теория относительности Эйнштейна. Нередко можно услышать, что физики всего лишь подмечают время от времени, где именно на этот раз математические концепции оказались хорошо применимыми к физическому поведению. Утверждают, соответственно, что физики, как правило, направляют свои интересы в сторону тех областей, где имеющиеся математические описания работают; таким образом, нет ничего удивительного в том, что математические и физические описания так хорошо друг с другом уживаются. Мне, впрочем, представляется, что авторы подобных заявлений, что называется, попадают пальцем в небо. Они просто никак не объясняют то фундаментальное единство, которое, как показывает, в частности, теория Эйнштейна, существует между математикой и устройством мироздания. Когда Эйнштейн разрабатывал свою теорию, никакой действительной необходимости в ней, с экспериментальной точки зрения, не было. Ньютоновская теория тяготения держалась уже почти 250 лет и достигла за это время потрясающей точности (погрешность порядка одной десятимиллионной — одно это является достаточно убедительным доказательством глубинной математической основы физической реальности). Да, в движении планеты Меркурий была замечена аномалия, однако это, разумеется, не послужило поводом для отказа от схемы Ньютона. И все же Эйнштейн счел, что можно добиться лучшего результата, если изменить саму основу теории тяготения. В первые годы после того, как Эйнштейн обнародовал теорию относительности, в поддержку ее можно было привести лишь несколько наблюдаемых эффектов, а преимущество над теорией Ньютона в точности было крайне незначительным. Теперь же, по прошествии 80 лет, общая точность теории относительности возросла в миллионы раз. Эйнштейн не просто «подметил» повторяющиеся особенности поведения физических объектов. Он обнаружил фундаментальную математическую субструктуру, реально существующую и до тех пор скрытую в глубинах мироздания. Более того, он искал вовсе не какие-то физические феномены, которые могли бы подойти под красивую теорию. Он искал и нашел точное математическое соотношение, заложенное в самой структуре пространства и времени, — наиболее фундаментальное из всех физических понятий.



В основе всех других успешных теорий элементарных физических процессов всегда лежит некая математическая структура, которая оказывается не только чрезвычайно точной, но и весьма хитроумной математически. (А чтобы читатель не подумал, что «ниспровержение» прежних физических представлений — например, теории Ньютона — каким-то образом эти представления обесценивает и лишает смысла, спешу уверить, что это ни в коем случае не так. Если прежние идеи были достаточно обоснованны — что можно сказать, например, о теориях Галилея или того же Ньютона, — то они и дальше остаются в добром здравии и находят в новой схеме свое место.) Кроме того, и сама математика, в своем стремлении как можно точнее описать поведение природных объектов, находит для себя немало полезного, порой неочевидного и неожиданного. И квантовая теория (тесные взаимоотношения которой с математикой — через посредство комплексных чисел — очевидны, надеюсь, даже из того краткого обзора предмета, что попал на эти страницы), и общая теория относительности, и электромагнитные уравнения Максвелла — все они дали весьма ощутимый толчок развитию математики. Причем это верно не только для относительно новых теорий,

что я перечислил. Не менее верно это и для теорий, куда более отдаленных от нас во времени, — например, для ньютоновской механики (давшей нам математический анализ) или древнегреческого анализа структуры пространства (которому мы обязаны самим понятием геометрии). Необычайная точность математики в описании физического поведения (например, точность квантовой электродинамики, достигающая одиннадцатого или даже двенадцатого знака после запятой) не раз удивляла ученых. Однако на этом загадки не заканчиваются. Концепции, скрывающиеся в физических процессах, обладают чрезвычайной глубиной, тонкостью и *математической плодотворностью*. Об этом люди зачастую и не подозревают — если, конечно, они не математики, вплотную занимающиеся соответствующей проблемой.

Следует особо подчеркнуть, что эта математическая плодотворность, дающая математикам ценный стимул в их работе, не является всего лишь следствием некоей математической моды (хотя и мода, надо признать, играет во всем этом свою роль). Идеи, которые были разработаны с единственной целью углубить наше понимание устройства физического мира, очень часто дают неожиданные и удивительно эффективные средства для решения других математических задач, которые *уже* какое-то время интенсивно и безуспешно пытаются решить другие люди совсем для других целей. В качестве одного из наиболее ярких недавних примеров можно привести найденное оксфордским математиком Саймоном Доналдсоном применение теорий типа Янга – Миллса (разработанных физиками в процессе отыскания математического объяснения взаимодействий между субатомными частицами) к исследованию четырехмерных многообразий[8], в результате чего были объяснены некоторые совершенно неожиданные их свойства, над которыми ученые бились в течение нескольких предыдущих лет. Что самое интересное, все эти математические средства (несмотря на то, что мы и не подозревали об их существовании, пока нас не посетило соответствующее озарение) вечно пребывают в безвременьи платоновского мира — неизменные истины, ожидающие своего открытия и открывающиеся лишь тем, кто обладает достаточным мастерством, проницательностью и упорством.

Надеюсь, мне удалось убедить читателя в существовании тесной и вполне реальной (хотя и все еще крайне загадочной) взаимосвязи между платоновским математическим миром и миром физических объектов. Надеюсь также, что само наличие такой взаимосвязи поможет скептикам отнестись к платоновскому миру именно как к «миру» несколько более серьезно, нежели они полагали для себя возможным прежде. Может быть, кто-то даже шагнет еще дальше, на что я рамках данного обсуждения не осмелился. Возможно, реальностью в платоновском смысле следует наделить и прочие абстрактные концепции, а не только математические. Сам Платон настаивал, что идеальные понятия «добра» и «красоты» реальны (см. § 8.3) ничуть не меньше, чем математические идеи. Лично у меня такая возможность никакого неприятия не вызывает, однако в моих размышлениях здесь она пока не играет сколько-нибудь серьезной роли. Я не уделил вопросам этики, морали и эстетики надлежащего внимания, однако это не повод для того, чтобы напрочь отказывать им в той же «реальности», какая досталась концепциям, которые рассмотрения удостоились. Безусловно, есть множество важных и разнообразных вопросов, которые следует изучить в этой связи, однако цели, что я ставил перед собой при написании этой конкретной книги, несколько уже[9].



Не уделил я *большого* внимания и собственно загадке (стрелка 1 на рис. 8.1) той непостижимой и абсолютной роли, что платоновский математический мир играет в физическом мире, — даже того, что получили другие две, о которых мы имеем еще меньшее представление. В первой части я обращался, по большей части, к вопросам, поднимаемым третьей стрелкой: загадкой нашего восприятия математического мира, т. е. выяснением природы процесса, посредством которого сознательное размышление способно «порождать», словно из ничего, те самые платоновские математические формы. (Как будто совершенные математические формы суть лишь тени наших несовершенных мыслей.) Такой взгляд на платоновский мир — как на продукт нашего сознания — весьма серьезно противоречит воззрениям самого Платона. Для Платона мир совершенных форм первичен, поскольку лежит вне времени и не зависит от человека. В истинно платоновском представлении мою третью стрелку на рис. 8.1 следует, очевидно, направить не вверх, а вниз: от мира совершенных форм к миру нашего сознания. Если же мы рассматриваем математический мир как продукт наших способов мышления, то это будет уже не платоновское представление, которого я здесь придерживаюсь, а самое настоящее *кантианство*.

Возможно, кому-то захочется аналогичным образом оспорить и направления остальных моих стрелок. Например, епископ Беркли, скорее всего, предпочел бы развернуть *вторую* стрелку, направить ее от ментального мира к миру физическому, поскольку, согласно его представлениям, «физическая реальность» есть лишь тень нашего ментального существования. Есть и такие (так называемые «номиналисты»), кто выступил бы за разворот *первой* стрелки, так как, по их мнению, мир математики является не более чем отражением аспектов мира физической реальности. Я сам, как явствует из этой книги, являюсь весьма решительным противником разворота первых двух стрелок; возможно, не менее очевидно и то, что я чувствую себя несколько неловко, будучи вынужден направить *третью* стрелку на рис. 8.1 в направлении, явно кантианском! Для меня мир совершенных форм первичен (как и для Платона) — существование этого мира является чуть ли не логической необходимостью, — оба же прочих мира суть его тени.

По причине такого расхождения во мнениях относительно того, какой из миров на рис. 8.1 следует считать первичным, а какие вторичными, я порекомендовал бы взглянуть на стрелки несколько иначе. Существенным качеством стрелок на рис. 8.1 является не столько их направление, сколько тот факт, что каждая представляет такое соответствие, при котором лишь *малая* область одного мира «порождает» *весь* следующий мир целиком. Что касается первой стрелки: мне много раз указывали на то, что огромная часть мира математики (если судить по результатам деятельности самих математиков) если и имеет какое-то отношение к действительному физическому поведению, то весьма незначительное. Получите: в основе структуры нашей физической Вселенной может лежать лишь крохотная часть платоновского мира. Аналогичным образом, вторая стрелка символизирует тот факт, что существование нашего ментального мира есть продукт очень малой части мира физического — той части, где имеются в точности те условия, что необходимы для возникновения сознания, как, например, в мозге человека. Точно так же третья стрелка захватывает весьма небольшую область мира ментальной активности, а именно ту, что «заведует» абсолютными и вневременными вопросами — в особенности, математической истиной. Наша с вами ментальная жизнь проходит, по большей части, совсем в других местах.



Есть нечто парадоксальное в этих соответствиях: каждый мир, похоже, «возникает» всего лишь из крохотной части того мира, что ему предшествует. На рис. 8.1 я постарался этот парадокс подчеркнуть. Впрочем, я рассматриваю стрелки не как утверждения о каких-то действительных «возникновениях», а просто как символы имеющихся соответствий, поскольку не хочу умножать предрассудки, и без того окружающие вопрос о том, какой из миров следует считать первичным, вторичным или третичным, если там вообще уместно такое «старшинство».

И все же полностью избежать предрассудков (или просто предвзятости) на рис. 8.1 мне не удалось. Если верить рисунку, то следует предположить, что *целый* мир отражается *частью* (причем малой) своего предшественника. Возможно, мои предрассудки ошибочны. Возможно, какие-то аспекты поведения физического мира невозможно описать в точных математических терминах; возможно, какая-то ментальная жизнь не связана неразрывно с физическими структурами (такими, как мозг); возможно также, что существуют математические истины, которые *принципиально недоступны* человеческому пониманию или интуиции. Для того, чтобы учесть все эти альтернативные возможности, рисунок 8.1 следует перерисовать таким образом, чтобы какие-то из миров (или все) охватывались стрелкой из предыдущего мира не полностью.

В первой части я большое внимание уделил некоторым следствиям из знаменитой теоремы Гёделя о неполноте. Кто-то из читателей, возможно, придерживается мнения, что теорема Гёделя как раз и утверждает, что в мире платоновских математических истин имеются области, принципиально недоступные человеческому пониманию или интуиции. Надеюсь, что мои доказательства ясно показали, что это не так[10]. Те математические предположения, что упоминаются в остроумном доказательстве Гёделя, человеку вполне доступны — при условии, что они построены в рамках математических (формальных) систем, которые уже приняты нами как достоверные средства оценки математической истинности. Из доказательства Гёделя отнюдь не следует, что существуют недоступные математические истины. Из него *следует* лишь, что человеческая интуиция не укладывается ни в рамки формальной аргументации, ни в рамки вычислительных процедур. Более того, из него недвусмысленно следует само существование платоновского математического мира. Математическая ис-

тина не определяется произвольным образом по правилам некоей «искусственной» формальной системы, но имеет абсолютный характер и находится вне любой такой системы устанавливаемых правил. Поддержка платоновского мировоззрения (в противовес формализму) была одной из важных причин, побудивших Гёделя взяться за работу. С другой стороны, рассуждения Гёделя могут служить иллюстрацией глубокой непостижимости нашего математического восприятия. Для того чтобы такое восприятие возникло, мы не просто «вычисляем»; тут на самом глубинном уровне задействовано что-то еще — что-то, что было бы невозможно без собственно осознания, которое, в конечном счете, и формирует мир восприятий.

Во второй части мы занимались в основном вопросами, имеющими отношение ко второй стрелке (хотя их адекватное рассмотрение невозможно без некоторых отсылок к стрелке первой), посредством которой плотный физический мир способен каким-то образом вызывать *теневой* феномен, называемый нами сознанием. Как же из таких, казалось бы, бесперспективных ингредиентов, как материя, пространство и время, возникает такой тонкий феномен, как сознание? До ответа мы так и не добрались, однако я надеюсь, что читатели смогли составить представление о загадочной природе как *самой* материи, так и пространства-времени, в рамках структуры которого оперируют теперь физические теории. Мы просто-напросто не располагаем достаточными знаниями ни о природе материи, ни о законах, которые этой материей управляют, — достаточными для того, чтобы понять, какая ее организация (в физическом мире) необходима, чтобы возникло осознающее себя существо. Более того, чем глубже мы исследуем природу материи, тем более эфемерной, таинственной и математической эта материя становится. Мы можем спросить: что же такое материя согласно лучшим теориям, которыми располагает на настоящий момент наука? Ответ мы получим математический, причем не в столько виде системы уравнений (хотя и уравнения тоже важны), сколько в виде тонких математических концепций, для одного лишь правильного понимания которых потребуется некоторое время.

Если общая теория относительности Эйнштейна показала, насколько могут измениться, приняв таинственный и математический вид, наши самые, казалось бы, незыблемые понятия о природе пространства и времени, то с концепцией *материи* аналогичную шутку сыграла квантовая механика. Глубокое потрясение испытали не только представления о материи, но и наше видение реальности вообще. Как может быть так, что одна лишь контрфактуальная *возможность* какого-либо события — т. е. что-то, чего в действительности *не* произошло, — оказывает вполне ощутимое воздействие на то, что в этой самой действительности *происходит*? При всей непостижимости проявлений квантовой механики в ней есть что-то такое, что по крайней мере *кажется* куда более близким (чем все, что может предложить классическая физика) к другой непостижимости, — той, за которой скрывается объяснение феномена ментальности в мире физической реальности. Я нисколько не сомневаюсь в том, что с появлением более глубоких теорий сознание наконец займет свое место в физическом мире и перестанет выглядеть на его фоне той «белой вороной», какой оно выглядит сегодня.



В §§ 7.7 и 8.6 я попытался ответить на вопрос, какие физические условия могут оказаться подходящими для возникновения феномена сознания. Я, однако, никоим образом не рассматриваю сознание исключительно как результат когерентного перемещения надлежащего количества вещества согласно правилам той или иной **OR**-теории квантово-классического интерфейса. Как я, надеюсь, достаточно ясно показал, все эти вещи всего лишь дают возможность расчистить в пределах современной физической картины мира место для невычислительных процессов. Подлинное сознание предполагает способность осознавать бесконечное разнообразие качественно различных вещей — зеленый цвет травы, запах цветов, пение птиц или мягкость меха, а также течение времени, радость, беспокойство, удивление или отношение к новой идее. Мы имеем идеалы, питаем надежды, выражаем намерения и усилием воли управляем множеством различных движений нашего тела, необходимых для реализации упомянутых намерений. Благодаря исследованиям в области нейроанатомии, неврологических нарушений, психиатрии и психологии, мы многое знаем о тонких взаимосвязях между физическими свойствами мозга и нашими ментальными состояниями. Все это мы, несомненно, вполне способны объяснить в терминах одной лишь физики критических объемов когерентного перемещения вещества. Однако без прорыва в новую физику мы так и останемся связаны смирительной рубашкой полностью вычислительной (или вычислительной вперемешку со случайной)

физики. Внутри мы не найдем научного объяснения ни интенциональности, ни субъективному опыту. Вырвавшись же из пут, мы получаем, по крайней мере, шанс когда-нибудь такое объяснение отыскать.

Многие, кто с этим согласится, добавят, что объяснения таким вещам не даст *никакая* научная картина. Тем, кто придерживается подобных взглядов, я могу лишь пожелать проявить немного терпения: подождем и посмотрим, как продвинется наука в будущем. Я думаю, что уже сейчас имеются некоторые указания (в загадочных процедурах квантовой механики) на то, что ментальные концепции стали ближе к нашим представлениям о физической вселенной, нежели прежде, — пусть и всего лишь *чуть* ближе. Я убежден, что с обнаружением необходимых *новых* физических принципов эти указания станут куда более отчетливыми. Науке еще есть куда развиваться; уж в этом-то сомневаться не приходится.

Более того, сама возможность понимания таких вещей человеком многое говорит о тех способностях, что дает нам сознание. Следует признать, что время от времени встречаются люди — например, Ньютон и Эйнштейн, Архимед и Галилей, Максвелл и Дирак; или Дарвин, Леонардо да Винчи, Рембрандт, Пикассо, Бах, Моцарт, Платон или те великие умы, что смогли породить такие шедевры, как «Илиада» или «Гамлет», — которые, по-видимому, наделены способностью «чувствовать» истину или красоту в значительно большей степени, нежели отпущено остальным. Однако единство с этой природной механикой потенциально присутствует во всех нас, проявляясь в способности к сознательному пониманию и ощущениям, на каком бы уровне эти процессы ни происходили. Каждый осознающий себя мозг сплетен из тончайших физических составляющих, неясным пока образом извлекающих сознание из фундаментальной структуры математически обусловленной Вселенной — с тем, чтобы мы, в свою очередь, смогли, вооружившись платоновским «пониманием», получить своего рода прямой доступ к первопричинам функционирования Вселенной на всевозможных уровнях.

Вопросы эти чрезвычайно глубоки и пока еще очень далеки от объяснения. Я утверждаю, что однозначных ответов мы не получим до тех пор, пока не поймем, как именно взаимодействуют между собой *все* три мира. Не получим мы ответов и в том случае, если будем пытаться разрешить каждый из вопросов отдельно от



остальных. Я говорил о трех мирах и трех загадках, связывающих их друг с другом. Разумеется, в действительности миров вовсе не три — мир всего *один*, и о его истинной природе мы все еще не имеем ни малейшего представления.

## Примечания

1. См., напр., [242].
2. Эту идею мне описал Жоэль де Роснэ.
3. «Слоны» (*Echo of the elephants*, BBC, январь 1993).
4. «Если не пойдут дожди» (*If the rains don't come*, BBC, сентябрь 1993).
5. «Грабеж среди бела дня» (*Daylight robbery*, BBC, август 1993).
6. Здесь можно поразмышлять на тему отсутствия (как правило) центриолей в нейронах (см. с. 557). Цитоскелеты клеток других типов, похоже, нуждаются в наличии центросом — с тем, чтобы те выполняли функции «управляющего центра» (необходимого для деления клетки), — цитоскелеты же нейронов, по всей вероятности, полагаются на власти более глобальные!
7. См. [257] и, напр., [38].
8. [96]; неплохое изложение вопроса для нематематиков имеется в [89] (гл. 10).
9. Объекты, которые разместились бы в таком расширенном платоновском мире, несколько напоминают те ментальные конструкции, что содержит попперовский «Мир 3»; см. [309]. Однако «Мир 3» не претендует ни на вневременное, независимое от нас существование, ни на то, чтобы служить основой для физической реальности. Соответственно, статус его существенно отличается от статуса того «платоновского мира», что рассматриваем мы с вами.
10. Во введении в свою книгу [270] Мостовски ясно показывает, что аргументы, подобные гёделевским, не имеют никакого отношения к вопросу о возможности существования *абсолютно* неразрешимых математических задач. На настоящий момент вопрос следует считать полностью открытым — нет ни доказательства, ни опровержения. Как и в случае с двумя другими стрелками, нам остается лишь верить или не верить.

# ЭПИЛОГ

Джессика с отцом вышли из пещеры. Снаружи было уже совсем темно и тихо, в прозрачном небе начали появляться звезды. Джессика повернулась к отцу.

— Знаешь, пап, вот я смотрю в небо, и мне все равно не верится, что Земля и *вправду* движется — и не только сама крутится вокруг оси, так еще и летит куда-то со скоростью сто тысяч километров в час, — хоть на самом деле я и знаю, что все это *должно* быть правдой.

Она замолчала и некоторое время просто стояла, глядя на звезды.

— Пап, расскажи мне о звездах...

# ЛИТЕРАТУРА


[1] Aharonov, Y., Albert, D. Z. (1981). Can we make sense out of the measurement process in relativistic quantum mechanics? *Phys. Rev.*, **D24**, 359–370.

[2] Aharonov, Y., Vaidman, L. (1990). Properties of a quantum system during the time interval between two measurements. *Phys. Rev.*, **A41**, 11.

[3] Aharonov, Y., Anandan, J., Vaidman, L. (1993). Meaning of the wave function. *Phys. Rev.*, **A47**, 4616–4626.

[4] Aharonov, Y., Bergmann, P. G., Liebowitz, J. L. (1964). Time symmetry in the quantum process of measurement. В сб. *Quantum theory and measurement* (ed. J. A. Wheeler, W. H. Zurek). Princeton University Press, 1983; первоначально в *Phys. Rev.*, **B134**, 1410–1416.

[5] Aharonov, Y., Albert, D. Z., Vaidman, L. (1986). Measurement process in relativistic quantum theory. *Phys. Rev.*, **D34**, 1805–1813.

[6] Albert, D. Z. (1983). On quantum-mechanical automata. *Phys. Lett.*, **98A (5, 6)**, 249–252.

[7] Albrecht-Buehler, G. (1981). Does the geometric design of centrioles imply their function? *Cell Motility*, **1**, 237–245.

[8] Albrecht-Buehler, G. (1985). Is the cytoplasm intelligent too? *Cell and Muscle Motility*, **6**, 1–21.

[9] Albrecht-Buehler, G. (1991). Surface extensions of 3T3 cells towards distant infrared light sources. *J. Cell Biol.*, **114**, 493–502.

[10] Anthony, M., Biggs, N. (1992). *Computational learning theory, an introduction*. Cambridge University Press.

[11] Applewhite, P.B. (1979). Learning in protozoa. В сб. *Biochemistry and physiology of protozoa*. Vol. 1 (ed. M. Levandowsky, S. H. Hunter), 341–355. Academic Press, New York.

[12] Arhem, P., Lindahl, B. I. B. (ed.) (1993). Neuroscience and the problem of consciousness: theoretical and empirical approaches. В сб. *Theoretical medicine*, **14**, Number 2. Kluwer Academic Publishers.

[13] Aspect, A., Grangier, P. (1986). Experiments on Einstein–Podolsky–Rosen-type correlations with pairs of visible photons. В сб. *Quantum concepts in space and time* (ed. R. Penrose, C. J. Isham). Oxford University Press.

[14] Aspect, A., Grangier, P., Roger, G. (1982). Experimental realization of Einstein–Podolsky–Rosen–Bohm *Gedanken-experiment*: a new violation of Bell's inequalities. *Phys. Rev. Lett.*, **48**, 91–94.

[15] Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.

[16] Bailey, T. N., Baston, R. J. (ed.) (1990). *Twistors in mathematics and physics*. London Mathematical Society Lecture Notes Series, 156. Cambridge University Press.

[17] Baylor, D. A., Lamb, T. D., Yau, K.-W. (1979). Responses of retinal rods to single photons. *J. Physiol.*, **288**, 613–634.

[18] Beck, F., Eccles, J. C. (1992). Quantum aspects of consciousness and the role of consciousness. *Proc. Nat. Acad. Sci.*, **89**, 11357–11361.

[19] Becks, K.-H., Hemker, A. (1992). An artificial intelligence approach to data analysis. В сб. *Proceedings of 1991 CERN School of Computing* (ed. C. Verkerk). CERN, Switzerland.

[20] Bell, J. S. (1964). On the Einstein Podolsky Rosen paradox. *Physics*, **1**, 195–200.

[21] Bell, J. S. (1966). On the problem of hidden variables in quantum theory. *Revs. Mod. Phys.*, **38**, 447–452.

[22] Bell, J. S. (1987). *Speakable and unspeakable in quantum mechanics*. Cambridge University Press.



[23] Bell, J. S. (1990). Against measurement. *Physics World*, **3**, 33–40.

[24] Benacerraf, P. (1967). God, the Devil and Gödel. *The Monist*, **51**, 9–32.

[25] Benioff, P. (1982). Quantum mechanical Hamiltonian models of Turing Machines. *J. Stat. Phys.*, **29**, 515–546.

[26] Bennett, C. H., Brassard, G., Breidbart, S., Wiesner, S. (1983). Quantum cryptography, or unforgettable subway tokens. В сб. *Advances in cryptography*. Plenum, New York.

[27] Bernard, C. (1875). *Leçons sur les anesthésiques et sur l'asphyxie*. J. B. Bailliere, Paris.

[28] Blakemore, C., Greenfield, S. (ed.) (1987). *Mind-waves: thoughts on intelligence, identity and consciousness*. Blackwell, Oxford.

[29] Blum, L., Shub, M., Smale, S. (1989). On a theory of computation and complexity over the real numbers: NP completeness, recursive functions and universal machines. *Bull. Amer. Math. Soc.*, **21**, 1–46.

[30] Bock, G. R., Marsh, J. (1993). *Experimental and theoretical studies of consciousness*. Wiley.

[31] Boden, M. (1977). *Artificial intelligence and natural man*. The Harvester Press, Hassocks.

[32] Boden, M. A. (1990). *The creative mind: myths and mechanisms*. Wiedenfeld and Nicolson, London.

[33] Bohm, D. (1952). A suggested interpretation of the quantum theory in terms of "hidden"variables, I and II. В сб. *Quantum theory and measurement* (ed. J. A. Wheeler, W. H. Zurek). Princeton University Press 1983. Первоначально в *Phys. Rev.*, **85**, 166–193.

[34] Bohm, D., Hiley, B. (1994). *The undivided universe*. Routledge, London.

[35] Boole, G. (1854). *An investigation of the laws of thought*. 1958, Dover, New York.

[36] Boolos, G. (1990). On seeing the truth of the Gödel sentence. *Behavioural and Brain Sciences*, **13 (4)**, 655.

[37] Bowie, G.L. (1982). Lucas' number is finally up. *J. of Philosophical Logic*, **11**, 279–285.

[38] Brady, M. (1993). Computational vision. В сб. *The simulation of human intelligence* (ed. D. Broadbent). Blackwell, Oxford.

[39] Braginsky, V.B. (1977). The detection of gravitational waves and quantum non-disturbative measurements. В сб. *Topics in theoretical and experimental gravitation physics* (ed. V. de Sabbata, J. Weber), 105. Plenum, London.

[40] Broadbent, D. (1993). Comparison with human experiments. В сб. *The simulation of human intelligence* (ed. D. Broadbent). Blackwell, Oxford.

[41] Brown, H.R. (1993). Bell's other theorem and its connection with nonlocality. Part I. В сб. *Bell's Theorem and the foundations of physics* (ed. A. Van der Merwe, F. Selleri). World Scientific, Singapore.

[42] Butterfield, J. (1990). Lucas revived? An undefended flank. *Behavioural and Brain Sciences*, **13 (4)**, 658.

[43] Castagnoli, G., Rasetti, M., Vincenti, A. (1992). Steady, simultaneous quantum computation: a paradigm for the investigation of nondeterministic and non-recursive computation. *Int. J. Mod. Phys. C*, **3**, 661–689.

[44] Caudill, M. (1992). *In our own image. Building an artificial person*. Oxford University Press.

[45] Chaitin, G.J. (1975). Randomness and mathematical proof. *Scientific American* (May 1975), 47.

[46] Chalmers, D.J. (1990). Computing the thinkable. *Behavioural and Brain Sciences*, **13 (4)**, 658.

[47] Chandrasekhar, S. (1987). *Truth and beauty. Aesthetics and motivations in science*. The University of Chicago Press.



[48] Chang, C.-L., Lee, R.C.-T. (1987). *Symbolic logic and mechanical theorem proving*, 2nd edn (1st edn 1973). Academic Press, New York.

[49] Chou, S.-C. (1988). *Mechanical geometry theorem proving*. Ridel.

[50] Christian, J.J. (1994). On definite events in a generally covariant quantum world. Unpublished preprint.

[51] Church, A. (1936). An unsolvable problem of elementary number theory. *Am. Jour. of Math.*, **58**, 345–363.

[52] Church, A. (1941). *The calculi of lambda-conversion*. Annals of Mathematics Studies, No. 6. Princeton University Press.

[53] Churchland, P.M. (1984). *Matter and consciousness*. Bradford Books, MIT Press, Cambridge, Massachusetts.

[54] Clauser, J.F., Horne, M.A. (1974). Experimental consequences of objective local theories. *Phys. Rev.*, **D10**, 526–535.

[55] Clauser, J.F., Horne, M.A., Shimony, A. (1978). Bell's theorem: experimental tests and implications. *Rpts. on Prog. in Phys.*, **41**, 1881–1927.

[56] Cohen, P.C. (1966). *Set theory and the continuum hypothesis*. Benjamin, Menlo Park, CA.

[57] Conrad, M. (1990). Molecular computing. В сб. *Advances in computers* (ed. M.C. Yovits), Vol. 31. Academic Press, London.

[58] Conrad, M. (1992). Molecular computing: the lock-key paradigm. *Computer* (November 1992), 11–20.

[59] Conrad, M. (1993). The fluctuon model of Force, Life, and computation: a constructive analysis. *Appl. Math. and Comp.*, **56**, 203–259.

[60] Cooke, 1988.

[61] Costa de Beauregard, O. (1989). В сб. *Bell's theorem, quantum theory, and conceptions of the universe* (ed. M. Kafatos). Kluwer, Dordrecht.

[62] Craik, K. (1943). *The nature of explanation.* Cambridge University Press.

[63] Crick, F. (1994). *The astonishing hypothesis. The scientific search for the soul.* Charles Scribner's Sons, New York, and Maxwell Macmillan International.

[64] Crick, F., Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences*, **2**, 263–275.

[65] Crick, F., Koch, C. (1992). The problem of consciousness. *Scientific American*, **267**, 110.

[66] Curl, R. F., Smalley, R. E. (1991). Fullerenes. *Scientific American*, **265**, No. 4, 32–41.

[67] Cutland N. J. (1980). *Computability. An introduction to recursive function theory.* Cambridge University Press.

[68] Davenport, H. (1952). *The higher arithmetic.* Hutchinson's University Library.

[69] Davies, P. C. W. (1974). *The physics of time asymmetry.* Surrey University Press, Belfast.

[70] Davies, P. C. W. (1984). *Quantum mechanics.* Routledge, London.

[71] Davis, M. (ed.) (1965). *The undecidable — basic papers on undecidable propositions, unsolvable problems and computable functions.* Raven Press, Hewlett, New York.

[72] Davis, M. (1978). What is a computation? В сб. *Mathematics today; twelve informal essays* (ed. L. A. Steen). Springer-Verlag, New York.

[73] Davis M. (1990). Is mathematical insight algorithmic? *Behavioural and Brain Sciences*, **13 (4)**, 659.

[74] Davis, M. (1993). How subtle is Gödel's theorem? *Behavioural and Brain Sciences*, **16**, 611–612.

[75] Davis, M., Hersch, R. (1975). Hilbert's tenth problem. *Scientific American* (Nov. 1973), 84.

[76] Davis, P. J., Hersch, R. (1982). *The mathematical experience.* Harvester Press.



[77] de Broglie, L. (1956). *Tentative d'intérpretation causale et nonlinéaire de la mécanique ondulatoire.* Gauthier-Villars, Paris.

[78] Deeke, L., Grötzinger, B., Kornhuber, H. H. (1976). Voluntary finger movements in man: cerebral potentials and theory. *Biol. Cybernetics*, **23**, 99.

[79] del Giudice, E., Doglia, S., Milani, M. (1983). Self-focusing and ponderomotive forces of coherent electric waves — a mechanism for cytoskeleton formation and dynamics. В сб. *Coherent excitations in biological systems* (ed. H. Fröhlich, F. Kremer). Springer-Verlag, Berlin.

[80] Dennett, D. (1990). Betting your life on an algorithm. *Behavioural and Brain Sciences*, **13 (4)**, 660.

[81] Dennett, D. C. (1991). *Consciousness explained.* Little, Brown and Company.

[82] d'Espagnat, B. (1989). *Conceptual foundations of quantum mechanics*, 2nd edn. Addison-Wesley, Reading, Massachusetts.

[83] Deutsch, D. (1985). Quantum theory, the Church–Turing principle and the universal quantum computer. *Proc. Roy. Soc. (Lond.)*, **A400**, 97–117.

[84] Deutsch, D. (1989). Quantum computational networks. *Proc. Roy. Soc. (Lond.)*, **A425**, 73–90.

[85] Deutsch, D. (1991). Quantum mechanics near closed time-like lines. *Phys. Rev.*, **D44**, 3197–3217.

[86] Deutsch, D. (1992). Quantum computation. *Phys. World*, **5**, 57–61.

[87] Deutsch, D., Ekert, A. (1993). Quantum communication moves into the unknown. *Phys. World*, **6**, 22–23.

[88] Deutsch, D., Jozsa, R. (1992). Rapid solution of problems by quantum computation. *Proc. R. Soc. Lond.*, **A439**, 553–558.

[89] Devlin, K. (1988). *Mathematics: the New Golden Age.* Penguin Books, London.

[90] DeWitt, B. S., Graham, R. D. (ed.)(1973). *The many-worlds interpretation of quantum mechanics*. Princeton University Press.

[91] Dicke, R. H. (1981). Interaction-free quantum measurements: a paradox? *Am. J. Phys.*, **49**, 925–930.

[92] Diósi, L. (1989). Models for universal reduction of macroscopic quantum fluctuations. *Phys. Rev.*, **A40**, 1165–1174.

[93] Diósi, L. (1992). Quantum measurement and gravity for each other. В сб. *Quantum chaos, quantum measurement*; NATO AS1 Series C. Math. Phys. Sci 357 (ed. P. Cvitanovic, I. C. Percival, A. Wirzba). Kluwer, Dordrecht.

[94] Dirac, P. A. M. (1947). *The principles of quantum mechanics*, 3rd edn. Oxford University Press.

[95] Dodd, A. (1991). Gödel, Penrose, and the possibility of AI. *Artificial Intelligence Review*, **5**.

[96] Donaldson, S. K. (1983). An application of gauge theory to four dimensional topology. *J. Diff. Geom.*, **18**, 279–315.

[97] Doyle, J. (1990). Perceptive questions about computation and cognition. *Behavioural and Brain Sciences*, **13 (4)**, 661.

[98] Dreyfus, H. L. (1972). *What computers can't do*. Harper and Row, New York.

[99] Dummett, M. (1973). *Frege: philosophy of language*. Duckworth, London.

[100] Dustin, P. (1984). *Microtubules*, 2nd revised edn. Springer-Verlag, Berlin.

[101] Dryl, S. (1974). Behaviour and motor responses in paramecium. В сб. *Paramecium — a current survey* (ed. W. J. Van Wagtendonk), 165–218. Elsevier, Amsterdam.

[102] Eccles, J. C. (1973). *The understanding of the brain*. McGraw-Hill, New York.

[103] Eccles, J. C. (1989). *Evolution of the brain: creation of the self*. Routledge, London.



[104] Eccles, J. C. (1992). Evolution of consciousness. *Proc. Natl. Acad. Sci.*, **89**, 7320–7324.

[105] Eccles, J. C. (1994). *How the self controls its brain*. Springer-Verlag, Berlin.

[106] Eckert, R., Randall, D., Augustine, G. (1988). *Animal physiology. Mechanisms and adaptations*, Chapter 11. Freeman, New York.

[107] Eckhorn, R., Bauer, R., Jordan, W., Brosch, M., Kruse, W., Munk, M., Reitboeck, H. J. (1988). Coherent oscillations: a mechanism of feature linking in the visual cortex? *Biol. Cybern.*, **60**, 121–130.

[108] Edelman, G. M. (1976). Surface modulation and cell recognition on cell growth. *Science*, **192**, 218–226.

[109] Edelman, G. M. (1987). *Neural Darwinism, the theory of neuronal group selection*. Basic Books, New York.

[110] Edelman, G. M. (1988). *Topobiology, an introduction to molecular embryology*. Basic Books, New York.

[111] Edelman, G. M. (1989). *The remembered present. A biological theory of consciousness*. Basic Books, New York.

[112] Edelman, G. M. (1992). *Bright air, brilliant fire: on the matter of the mind*. Allen Lane, The Penguin Press, London.

[113] Einstein, A., Podolsky, P., Rosen, N. (1935). Can quantum-mechanical description of physical reality be considered complete? В сб. *Quantum theory and measurement* (ed. J. A. Wheeler, W. H. Zurek). Princeton University Press, 1983. Первоначально в *Phys. Rev.*, **47**, 777–780.

[114] Elitzur, A. C., Vaidman, L. (1993). Quantum-mechanical interaction-free measurements. *Found. of Phys.*, **23**, 987–997.

[115] Elkies, Noam G. (1988). On $A^4 + B^4 + C^4 = D^4$. *Maths. of Computation*, **51**, (No. 184), 825–835.

[116] Everett, H. (1957). "Relative State"formulation of quantum mechanics. В сб. *Quantum theory and measurement* (ed. J. A. Wheeler, W. H. Zurek). Princeton University Press,

1983; первоначально в *Rev. of Modern Physics*, **29**, 454–462.

[117] Feferman, S. (1988). Turing in the Land of O(z). В сб. *The universal Turing machine: a half-century survey* (ed. R. Herken). Kammerer and Unverzagt, Hamburg.

[118] Feynman, R.P. (1948). Space-time approach to non-relativistic quantum mechanics. *Revs. Mod. Phys.*, **20**, 367–387.

[119] Feynman, R.P. (1982). Simulating physics with computers. *Int. J. Theor. Phys.*, **21 (6/7)**, 467–488.

[120] Feynman, R.P. (1985). Quantum mechanical computers. *Optics News*, Feb., 11–20.

[121] Feynman, R.P. (1986). Quantum mechanical computers. *Foundations of Physics*, **16 (6)**, 507–531.

[122] Fodor, J.A. (1983). *The modularity of mind.* MIT Press, Cambridge, Massachusetts.

[123] Franks, N.P., Lieb, W.R. (1982). Molecular mechanics of general anaesthesia. *Nature*, **300**, 487–493.

[124] Freedman, D.H. (1994). *Brainmakers*. Simon and Schuster, New York.

[125] Freedman, S.J., Clauser, J.F. (1972). Experimental test of local hidden-variable theories. В сб. *Quantum theory and measurement* (ed. J.A. Wheeler, W.H. Zurek). Princeton University Press, 1983; первоначально в *Phys. Rev. Lett.*, **28**, 938–941.

[126] Frege, G. (1893). *Grundgesetze der Arithmetik, begriffsschriftlich abelgeleitet*, Vol. 1. H. Pohle, Jena.

[127] Frege, G. (1964). *The basic laws of arithmetic*, translated and edited with an introduction by Montgomery Firth. University of California Press, Berkeley.

[128] French, J.W. (1940). Trial and error learning in paramecium. *J. Exp. Psychol.*, **26**, 609–613.

[129] Fröhlich, H. (1968). Long-range coherence and energy storage in biological systems. *Int. Jour. of Quantum. Chem.*, **II**, 641–649.



[130] Fröhlich, H. (1970). Long range coherence and the actions of enzymes. *Nature*, **228**, 1093.

[131] Fröhlich, H. (1975). The extraordinary dielectric properties of biological materials and the action of enzymes, *Proc. Natl. Acad. Sci.*, **72 (11)**, 4211–4215.

[132] Fröhlich, H. (1984). General theory of coherent excitations on biological systems. В сб. *Nonlinear electrodynamics in biological systems* (ed. W.R. Adey, A.F. Lawrence). Plenum Press, New York.

[133] Fröhlich, H. (1986). Coherent excitations in active biological systems. В сб. *Modern bioelectrochemistry* (ed. F. Gutmann, H. Keyzer). Plenum Press, New York.

[134] Fukui, K., Asai, H. (1976). Spiral motion of paramecium caudatum in small capillary glass tube. *J. Protozool.*, **23**, 559–563.

[135] Gandy, R. (1988). The confluence of ideas in 1936. В сб. *The universal Turing machine: a half-century survey* (ed. R. Herken). Kammerer and Unverzagt, Hamburg.

[136] Gardner, M. (1965). *Mathematical magic show.* Alfred Knopf, New York; Random House, Toronto.

[137] Gardner, M. (1970). Mathematical games: the fantastic combinations of John Conway's new solitaire game "Life". *Scientific American*, **223**, 120–123.

[138] Gardner, M. (1989). *Penrose tiles to trapdoor ciphers.* Freeman, New York.

[139] Gelber, B. (1958). Retention in paramecium aurelia. *J. Comp. Physiol. Psych.*, **51**, 110–115.

[140] Gelernter, D. (1994). *The muse in the machine.* The Free Press, Macmillan Inc., New York; Collier Macmillan, London.

[141] Gell-Mann, M., Hartle, J.B. (1993). Classical equations for quantum systems. *Phys. Rev.*, **D47**, 3345–3382.

[142] Gernoth, K.A., Clark, J.W., Prater, J.S., Bohr, H. (1993). Neural network models of nuclear systematics. *Phys. Lett.*, **B300**, 1–7.

[143] Geroch, R. (1984). The Everett interpretation. *Nous*, **4** (специальный выпуск, посвященный основным принципам квантовой механики), 617–633.

[144] Geroch, R., Hartle, J. B. (1986). Computability and physical theories. *Found. Phys.*, **16**, 533.

[145] Ghirardi, G. C., Rimini, A., Weber, T. (1980). A general argument against superluminal transmission through the quantum mechanical measurement process. *Lett. Nuovo Cim.*, **27**, 293–298.

[146] Ghirardi, G. C., Rimini, A., Weber, T. (1986). Unified dynamics for microscopic and macroscopic systems. *Phys. Rev.*, **D34**, 470.

[147] Ghirardi, G. C., Grassi, R., Rimini, A. (1990). Continuous-spontaneous-reduction model involving gravity. *Phys. Rev.*, **A42**, 1057–1064.

[148] Ghirardi, G. C., Grassi, R., Pearle, P. (1990). Relativistic dynamical reduction models: general framework and examples. *Foundations of Physics*, **20**, 1271–1316.

[149] Ghirardi, G. C., Grassi, R., Pearle, P. (1992). Comment on "Explicit collapse and superluminal signals". *Phys. Lett.*, **A166**, 435–438.

[150] Ghirardi, G. C., Grassi, R., Pearle, P. (1993). Negotiating the tricky border between quantum and classical. *Physics Today*, **46**, 13.

[151] Gisin, N. (1989). Stochastic quantum dynamics and relativity. *Helv. Phys. Acta*, **62**, 363–371.

[152] Gisin, N., Percival, I. C. (1993). Stochastic wave equations versus parallel world components. *Phys. Lett.*, **A175**, 144–145.

[153] Gleick, J. (1987). *Chaos. Making a new science*. Penguin Books.

[154] Glymour, C., Kelly, K. (1990). Why you'll never know whether Roger Penrose is a computer. *Behavioural and Brain Sciences*, **13 (4)**, 666.



[155] Gödel, K. (1931). Über formal unentscheidbare Sätze per Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, **38**, 173–198.

[156] Gödel, K. (1940). *The consistency of the axiom of choice and of the generalized continuum-hypothesis with the axioms of set theory*. Princeton University Press, Oxford University Press.

[157] Gödel, K. (1949). An example of a new type of cosmological solution of Einstein's field equations of gravitation. *Rev. of Mod. Phys.*, **21**, 447.

[158] Gödel, K. (1986). *Kurt Gödel, collected works*, Vol. I (publications 1929–1936) (ed. S. Feferman et al.). Oxford University Press.

[159] Gödel, K. (1990). *Kurt Gödel, collected works*, Vol. II (publications 1938–1974) (ed. S. Feferman et al.). Oxford University Press.

[160] Gödel, K. (1995). *Kurt Gödel, collected works*, Vol. III (ed. S. Feferman et al.). Oxford University Press.

[161] Golomb, S. W. (1965). *Polyominoes*. Scribner and Sons.

[162] Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. *Advances in Computers*, **6**, 31–88.

[163] Good, I. J. (1967). Human and machine logic. *Brit. J. Philos. Sci.*, **18**, 144–147.

[164] Good, I. J. (1969). Gödel's theorem is a red herring. *Brit. J. Philos. Sci.*, **18**, 359–373.

[165] Graham, R. L., Rothschild, B. L. (1971). Ramsey's theorem for $n$-parameter sets. *Trans. Am. Math. Soc.*, **59**, 290.

[166] Grant, P. M. (1994). Another December revolution? *Nature*, **367**, 16.

[167] Gray, C. M., Singer, W. (1989). Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proc. Natl. Acad. Sci. USA*, **86**, 1689–1702.

[168] Grangier, P., Roger, G., Aspect, A. (1986). Experimental evidence for a photon anticorrelation effect on a beam splitter: a new light on single-photon interferences. *Europhysics Letters*, **1**, 173–179.

[169] Green, D. G., Bossomaier, T. (ed.) (1993). *Complex systems: from biology to computation*. IOS Press.

[170] Greenberger, D. M., Horne, M. A., Zeilinger, A. (1989). Going beyond Bell's theorem. В сб. *Bell's theorem, quantum theory, and conceptions of the universe* (ed. M. Kafatos), 73–76. Kluwer Academic, Dordrecht, The Netherlands.

[171] Greenberger, D. M., Horne, M. A., Shimony, A., Zeilinger, A. (1990). Bell's theorem without inequalities. *Am. J. Phys.*, **58**, 1131–1143.

[172] Gregory, R. L. (1981). *Mind in science: a history of explanations in psychology and physics*. Weidenfeld and Nicholson Ltd. (также Penguin, 1984).

[173] Grey Walter, W. (1953). *The living brain*. Gerald Duckworth and Co. Ltd.

[174] Griffiths, R. (1984). Consistent histories and the interpretation of quantum mechanics. *J. Stat. Phys.*, **36**, 219.

[175] Grossberg, S. (ed.) (1987). *The adaptive brain I: Cognition, learning, reinforcement and rhythm* и *The adaptive brain II: Vision, speech, language and motor control*. North-Holland, Amsterdam.

[176] Grünbaum, B., Shephard, G. C. (1987). *Tilings and Patterns*. Freeman, New York.

[177] Grundler, W., Keilmann, F. (1983). Sharp resonances in yeast growth proved nonthermal sensitivity to microwaves. *Phys. Rev. Letts.*, **51**, 1214–1216.

[178] Guccione, S. (1993). Mind the truth: Penrose's new step in the Gödelian argument. *Behavioural and Brain Sciences*, **16**, 612–613.

[179] Haag, R. (1992). *Local quantum physics: fields, particles, algebras*. Springer-Verlag, Berlin.



[180] Hadamard, J. (1945). *The psychology of invention in the mathematical field*. Princeton University Press.

[181] Hallett, M. (1984). *Cantorian set theory and limitation of size*. Clarendon Press, Oxford.

[182] Hameroff, S. R. (1974). Chi: a neural hologram? *Am. J. Clin. Med.*, **2 (2)**, 163–170.

[183] Hameroff, S. R. (1987). *Ultimate computing. Biomolecular consciousness and nano-technology*. North-Holland, Amsterdam.

[184] Hameroff, S. R., Watt, R. C. (1982). Information in processing in microtubules. *J. Theor. Biol.*, **98**, 549–561.

[185] Hameroff, S. R., Watt, R. C. (1983). Do anesthetics act by altering electron mobility? *Anesth. Analg.*, **62**, 936–940.

[186] Hameroff, S. R., Rasmussen, S., Mansson, B. (1988). Molecular automata in microtubles: basic computational logic of the living state? В сб. *Artificial Life, SFI studies in the sciences of complexity* (ed. C. Langton). Addison-Wesley, New York.

[187] Hanbury Brown, R., Twiss, R. Q. (1954). A new type of interferometer for use in radio astronomy. *Phil. Mag.*, **45**, 663–682.

[188] Hanbury Brown, R., Twiss, R. Q. (1956). The question of correlation between photons in coherent beams of light. *Nature*, **177**, 27–29.

[189] Harel, D. (1987). *Algorithmics. The spirit of computing*. Addison-Wesley, New York.

[190] Hawking, S. W. (1975). Particle creation by Black Holes. *Commun. Math. Phys.*, **43**, 199–220.

[191] Hawking, S. W. (1982). Unpredictability of quantum gravity. *Commun. Math. Phys.*, **87**, 395–415.

[192] Hawking, S. W., Israel, W. (ed.) (1987). *300 years of gravitation*. Cambridge University Press.

[193] Hebb, D. O. (1949). *The organization of behaviour*. Wiley, New York.

[194] Hecht, S., Shlaer, S., Pirenne, M. H. (1941). Energy, quanta and vision. *Journal of General Physiology*, **25**, 821–840.

[195] Herbert, N. (1993). *Elemental mind. Human consciousness and the new physics*. Dutton Books, Penguin Publishing.

[196] Heyting, A. (1956). *Intuitionism: an introduction*. North-Holland, Amsterdam.

[197] Heywood, P., Redhead, M. L. G. (1983). Nonlocality and the Kochen–Specker Paradox. *Found. Phys.*, **13**, 481–499.

[198] Hodges, A. P. (1983). *Alan Turing: the enigma*. Burnett Books and Hutchinson, London; Simon and Schuster, New York.

[199] Hodgkin, D., Houston, A. I. (1990). Selecting for the con in consciousness. *Behavioural and Brain Sciences*, **13 (4)**, 668.

[200] Hodgson, D. (1991). *Mind matters: consciousness and choice in a quantum world*. Clarendon Press, Oxford.

[201] Hofstadter, D. R. (1979). *Gödel, Escher, Bach: an eternal golden braid*. Harvester Press, Hassocks, Essex.

[202] Hofstadter, D. R. (1981). A conversation with Einstein's brain. В сб. *The mind's I* (ed. D. R. Hofstadter, D. Dennett). Basic Books; Penguin, Harmondsworth, Middlesex.

[203] Hofstadter, D. R., Dennett, D. C. (ed.) (1981). *The mind's I*. Basic Books; Penguin, Harmondsworth, Middlesex.

[204] Home, D. (1994). A proposed new test of collapse-induced quantum nonlocality. Preprint.

[205] Home, D., Nair, R. (1994). Wave function collapse as a nonlocal quantum effect. *Phys. Lett.*, **A187**, 224–226.

[206] Home, D., Selleri, F. (1991). Bell's Theorem and the EPR Paradox. *Rivista del Nuovo Cimento*, **14**, N. 9.

[207] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci.*, **79**, 2554–2558.



[208] Hsu, F.-H., Anantharaman, T., Campbell, M., Nowatzyk, A. (1990). A grandmaster chess machine. *Scientific American*, **263**.

[209] Huggett, S. A., Tod, K. P. (1985). *An introduction to twistor theory*. London Math. Soc. student texts. Cambridge University Press.

[210] Hughston, L. P., Jozsa, R., Wootters, W. K. (1993). A complete classification of quantum ensembles having a given density matrix. *Phys. Letters*, **A183**, 14–18.

[211] Isham, C. J. (1989). Quantum gravity. В сб. *The new physics* (ed. P. C. W. Davies), 70–93. Cambridge University Press.

[212] Isham, C. J. (1994). Prima facie questions in quantum gravity. В сб. *Canonical relativity: classical and quantum* (ed. J. Ehlers, H. Friedrich). Springer-Veriag, Berlin.

[213] Jibu, M., Hagan, S., Pribram, K., Hameroff, S. R., Yasue, K. (1994). Quantum optical coherence in cytoskeletal microtubules: implications for brain function. *Bio. Systems* (готовится к печати).

[214] Johnson-Laird, P. N. (1983). *Mental models*. Cambridge University Press.

[215] Johnson-Laird, P. N. (1987). How could consciousness arise from the computations of the brain? В сб. *Mindwaves: thoughts on intelligence, identity and consciousness* (ed. C. Blakemore, S. Greenfield). Blackwell, Oxford.

[216] Károlyházy, F. (1966). Gravitation and quantum mechanics of macroscopic bodies. *Nuo. Cim. A*, **42**, 390–402.

[217] Károlyházy, F. (1974). Gravitation and quantum mechanics of macroscopic bodies. *Magyar Fizikai Polyoirat*, **12**, 24.

[218] Károlyházy, F., Frenkel, A., Lukacs, B. (1986). On the possible role of gravity on the reduction of the wave function. В сб. *Quantum concepts in space and time* (ed. R. Penrose, C. J. Isham). Oxford University Press.

[219] Kasumov, A. Y., Kislov, N. A., Khodos, I. I. (1993). Can the observed vibration of a cantilever of supersmall mass

be explained by quantum theory? *Microsc. Microanal. Microstruct.*, **4**, 401–406.

[220] Kentridge, R. W. (1990). Parallelism and patterns of thought. *Behavioural and Brain Sciences*, **13 (4)**, 670.

[221] Khalfa, J. (ed.) (1994). *What is intelligence? The Darwin College lectures*. Cambridge University Press.

[222] Klarner, D. A. (1981). My life among the Polyominoes. В сб. *The mathematical gardner* (ed. D. A. Klarner). Prindle, Weber and Schmidt, Boston, MA; Wadsworth Int., Belmont, CA.

[223] Kleene, S. C. (1952). *Introduction to metamathematics*. North-Holland, Amsterdam, van Nostrand, New York.

[224] Klein, M. V., Furtak, T. E. (1986). *Optics*, 2nd edn. Wiley, New York.

[225] Kochen, S., Specker, E. P. (1967). The problem of hidden variables in quantum mechanics. *J. Math. Mech.*, **17**, 59–88.

[226] Kohonen, T. (1984). *Self-organization and associative memory*. Springer-Verlag, New York.

[227] Komar, A. B. (1969). Qualitative features of quantized gravitation. *Int. J. Theor. Phys.*, **2**, 157–160.

[228] Koruga, D. (1974). Microtubule screw symmetry: packing of spheres as a latent bioinformation code. *Ann. NY Acad. Sci.*, **466**, 953–955.

[229] Koruga, D., Hameroff, S., Withers, J., Loutfy, R., Sundareshan, M. (1993). *Fullerene $C_{60}$. History, physics, nanobiology, nanotechnology*. North-Holland, Amsterdam.

[230] Kosko, B. (1994). *Fuzzy thinking: the new science of fuzzy logic*. Harper Collins, London.

[231] Kreisel, G. (1960). Ordinal logics and the characterization of informal concepts of proof. *Proc. of the Internal. Cong. of Mathematics, Aug. 1958*. Cambridge University Press.

[232] Kreisel, G. (1967). Informal rigour and completeness proofs. В сб. *Problems in the philosophy of mathematics* (ed. I. Lakatos), 138–186. North-Holland, Amsterdam.



[233] Laguës, M., Xiao Ming Xie, Tebbji, H., Xiang Zhen Xu, Mairet, V., Hatterer, C., et al. (1993). Evidence suggesting superconductivity at 250 K in a sequentially deposited cuprate film. *Science*, **262**, 1850–1851.

[234] Lander, L. J., Parkin, T. R. (1966). Counterexample to Euler's conjecture on sums of like powers. *Bull. Amer. Math. Soc.*, **72**, 1079.

[235] Leggett, A. J. (1984). Schrödinger's cat and her laboratory cousins. *Contemp. Phys.*, **25 (6)**, 583.

[236] Lewis, D. (1969). Lucas against mechanism. *Philosophy*, **44**, 231–233.

[237] Lewis, D. (1989). Lucas against mechanism II. *Can. J. Philos.*, **9**, 373–376.

[238] Libet, B. (1990). Cerebral processes that distinguish conscious experience from unconscious mental functions. В сб. *The principles of design and operation of the brain* (ed. J. C. Eccles, O. D. Creutzfeldt), Experimental Brain research series 21, 185–205. Springer-Verlag, Berlin.

[239] Libet, B. (1992). The neural time-factor in perception, volition and free will. *Revue de Metaphysique et de Morale*, **2**, 255–272.

[240] Libet, B., Wright, E. W. Jr., Feinstein, B., Pearl, D. K. (1979). Subjective referral of the timing for a conscious sensory experience. *Brain*, **102**, 193–224.

[241] Linden, E. (1993). Can animals think? *Time Magazine* (March), 13.

[242] Lisboa, P. G. J. (ed.) (1992). *Neural networks: current applications*. Chapman Hall, London.

[243] Lockwood, M. (1989). *Mind, brain and the quantum*. Blackwell, Oxford.

[244] Longair, M. S. (1993). Modern cosmology — a critical assessment. *Q. J. R. Astr. Soc.*, **34**, 157–199.

[245] Longuet-Higgins, H. C. (1987). *Mental processes: studies in cognitive science, Part II*. MIT Press, Cambridge, Massachusetts.

[246] Lucas, J.R. (1961). Minds, machines and Gödel. *Philosophy*, **36**, 120–124; также в Alan Ross Anderson (ed.) (1964) *Minds and Machines*. Englewood Cliffs.

[247] Lucas, J.R. (1970). *The freedom of the will*. Oxford University Press.

[248] McCarthy, J. (1979). Ascribing mental qualities to machines. В сб. *Philosophical perspectives in artificial intelligence* (ed. M. Ringle). Humanities Press, New York.

[249] McCulloch, W.S., Pitts, W.H. (1943). A logical calculus of the idea immanent in nervous activity. *Bull. Math. Biophys.*, **5**, 115–133. (Также в McCulloch, W.S., *Embodiments of mind*, MIT Press, 1965.)

[250] McDermott, D. (1990). Computation and consciousness. *Behavioural and Brain Sciences*, **13 (4)**, 676.

[251] MacLennan, B. (1990). The discomforts of dualism. *Behavioural and Brain Sciences*, **13 (4)**, 673.

[252] Majorana, E. (1932). Atomi orientati in campo magnetico variabile. *Nuovo Cimento*, **9**, 43–50.

[253] Manaster-Ramer, A., Savitch, W.J., Zadrozny, W. (1990). Gödel redux. *Behavioural and Brain Sciences*, **13 (4)**, 675.

[254] Mandelkow, E.-M., Mandelkow, F. (1994). Microtubule structure. *Curr. Opinions Structural Biology*, **4**, 171–179.

[255] Margulis, L. (1975). *Origins of eukaryotic cells*. Yale University Press, New Haven, CT.

[256] Markov, A.A. (1958). The insolubility of the problem of homeomorphy. *Dokl. Akad. Nauk. SSSR*, **121**, 218–220.

[257] Marr, D.E. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. Freeman, San Francisco.

[258] Marshall, I.N. (1989). Consciousness and Bose–Einstein condensates. *New Ideas in Psychology*, **7**.

[259] Mermin, D. (1985). Is the moon there when nobody looks? Reality and the quantum theory. *Physics Today*, **38**, 38–47.



[260] Mermin, D. (1990). Simple unified form of the major no-hidden-variables theorems. *Phys. Rev. Lett.*, **65**, 3373–3376.

[261] Michie, D., Johnston, R. (1984). *The creative computer. Machine intelligence and human knowledge*. Viking Penguin.

[262] Minsky, M. (1968). Matter, mind and models. В сб. *Semantic information processing* (ed. M. Minsky). MIT Press, Cambridge, Massachusetts.

[263] Minsky, M. (1986). *The society of mind*. Simon and Schuster, New York.

[264] Minsky, M., Papert, S. (1972). *Perceptrons: an introduction to computational geometry*. MIT Press, Cambridge, Massachusetts.

[265] Misner, C.W., Thorne, K.S., Wheeler, J.A. (1973). *Gravitation*. Freeman, New York.

[266] Moore, A.W. (1990). *The infinite*. Routledge, London.

[267] Moravec, H. (1988). *Mind children: the future of robot and human intelligence*. Harvard University Press, Cambridge, Massachusetts.

[268] Moravec, H. (1994). *The Age of Mind: transcending the human condition through robots*. Готовится к печати.

[269] Mortensen, C. (1990). The powers of machines and minds. *Behavioural and Brain Sciences*, **13 (4)**, 678.

[270] Mostowski, A. (1957). *Sentences undecidable in formalized arithmetic: an exposition of the theory of Kurt Gödel*. North-Holland, Amsterdam.

[271] Nagel, E., Newman, J.R. (1958). *Gödel's proof*. Routledge and Kegan Paul.

[272] Newell, A., Simon, H.A. (1976). Computer science as empirical enquiry: symbols and search. *Communications of the ACM*, **19**, 113–126.

[273] Newell, A., Young, R., Polk, T. (1993). The approach through symbols. В сб. *The simulation of human intelligence* (ed. D. Broadbent). Blackwell, Oxford.

[274] Newton, I. (1687). *Philosophiae Naturalis Principia Mathematica*. Репринт: Cambridge University Press.

[275] Newton, I. (1730). *Opticks*. 1952, Dover, New York.

[276] Oakley, D. A. (ed.) (1985). *Brain and mind*. Methuen, London.

[277] Obermayer, K., Teich, W. G., Mahler, G. (1988). Structural basis of multistationary quantum systems. I. Effective single-particle dynamics. *Phys. Rev.*, **B37**, 8096–8110.

[278] Obermayer, K., Teich, W. G., Mahler, G. (1988). Structural basis of multistationary quantum systems. II. Effective few-particle dynamics. *Phys. Rev.*, **B37**, 8111–8121.

[279] Omnès, R. (1992). Consistent interpretations of quantum mechanics. *Rev. Mod. Phys.*, **64**, 339–382.

[280] Pais, A. (1991). *Niels Bohr's times*. Clarendon Press, Oxford.

[281] Pauling L. (1964). The hydrate microcrystal theory of general anesthesia. *Anesth. Analg.*, **43**, 1.

[282] Paz, J. P., Zurek, W. H. (1993). Environment-induced decoherence, classicality and consistency of quantum histories. *Phys. Rev.*, **D48 (6)**, 2728–2738.

[283] Paz, J. P., Habib, S., Zurek, W. H. (1993). Reduction of the wave packet: preferred observable and decoherence time scale. *Phys. Rev.*, **D47 (2)**, 3rd Series, 488–501.

[284] Pearle, P. (1976). Reduction of the state-vector by a nonlinear Schrödinger equation. *Phys. Rev.*, **D13**, 857–868.

[285] Pearle, P. (1989). Combining stochastic dynamical state-vector reduction with spontaneous localization. *Phys. Rev.*, **A39**, 2277–2289.

[286] Pearle, P. (1992). Relativistic model state-vector reduction. В сб. *Quantum chaos — quantum measurement*, NATO Adv. Sci. Inst. Ser. C. Math. Phys. Sci. 358 (Copenhagen 1991). Kluwer, Dordrecht.

[287] Peat, F. D. (1988). *Superstrings and the search for the theory of everything*. Contemporary Books, Chicago.



[288] Penrose, O. (1970). *Foundations of statistical mechanics: a deductive treatment*. Pergamon, Oxford.

[289] Penrose, O., Onsager, L. (1956). Bose–Einstein condensation and liquid helium. *Phys. Rev.*, **104**, 576–584.

[290] Penrose, R. (1980). On Schwarzschild causality — a problem for "Lorentz covariant" general relativity. В сб. *Essays in general relativity* (A. Taub Festschrift) (ed. F. J. Tipler), 1–12. Academic Press, New York.

[291] Penrose, R. (1987). Newton, quantum theory and reality. В сб. *300 Years of gravity* (ed. S. W. Hawking, W. Israel). Cambridge University Press.

[292] Penrose, R. (1990). Author's response, *Behavioural and Brain Sciences*, **13 (4)**, 692.

[293] Penrose, R. (1991). The mass of the classical vacuum. В сб. *The philosophy of vacuum* (ed. S. Saunders, H. R. Brown). Clarendon Press, Oxford.

[294] Penrose, R. (1991). Response to Tony Dodd's "Gödel, Penrose, and the possibility of AI". *Artificial Intelligence Review*, **5**, 235.

[295] Penrose, R. (1993). Gravity and quantum mechanics. В сб. *General relativity and gravitation 1992. Proceedings of the Thirteenth International Conference on General Relativity and Gravitation held at Cordoba, Argentina 28 June–4 July 1992. Part 1: Plenary lectures* (ed. R. J. Gleiser, C. N. Kozameh, O. M. Moreschi). Institute of Physics Publications, Bristol.

[296] Penrose, R. (1993). Quantum non-locality and complex reality. В сб. *The Renaissance of general relativity* (in honour of D. W. Sciama) (ed. G. Ellis, A. Lanza, J. Miller). Cambridge University Press.

[297] Penrose, R. (1993). Setting the scene: the claim and the issues. В сб. *The simulation of human intelligence* (ed. D. Broadbent). Blackwell, Oxford.

[298] Penrose, R. (1993). An emperor still without mind. *Behavioural and Brain Sciences*, **16**, 616–622.

[299] Penrose, R. (1994). On Bell non-locality without probabilities: some curious geometry. В сб. *Quantum reflections* (in honour of J. S. Bell) (ed. J. Ellis, A. Amati). Cambridge University Press.

[300] Penrose, R. (1994). Non-locality and objectivity in quantum state reduction. В сб. *Fundamental aspects of quantum theory* (ed. J. Anandan, J. L. Safko). World Scientific, Singapore.

[301] Penrose, R., Rindler, W. (1984). *Spinors and space-time*, Vol. 1: *Two-spinor calculus and relative fields*. Cambridge University Press.

[302] Penrose, R., Rindler, W. (1986). *Spinors and space-time*, Vol. 2: *Spinor and twistor methods in space-time geometry*. Cambridge University Press.

[303] Percival, I. C. (1994). Primary state diffusion. *Proc. R. Soc. Lond.*, **A** (статья отправлена в журнал).

[304] Peres, A. (1985). Reversible logic and quantum computers. *Phys. Rev.*, **A32 (6)**, 3266–3276.

[305] Peres, A. (1990). Incompatible results of quantum measurements. *Phys. Lett.*, **A151**, 107–108.

[306] Peres, A. (1991). Two simple proofs of the Kochen–Specker theorem. *J. Phys. A: Math. Gen.*, **24**, L175–L178.

[307] Perlis, D. (1990). The emperor's old hat. *Behavioural and Brain Sciences*, **13 (4)**, 680.

[308] Planck, M. (1906). *The theory of heat radiation* (пер. на англ.: M. Masius, основана на лекциях, прочитанных в Берлине в 1906/1907 годах). 1959, Dover, New York.

[309] Popper, K. R., Eccles, J. R. (1977). *The self and its brain*. Springer International.

[310] Post, E. L. (1936). Finite combinatory processes-formulation I, *Jour. Symbolic Logic*, **1**, 103–105.

[311] Poundstone, W. (1985). *The recursive universe: cosmic complexity and the limits of scientific knowledge*. Oxford University Press.



[312] Pour-El, M. B. (1974). Abstract computability and its relation to the general purpose analog computer. (Some connections between logic, differential equations and analog computers.) Trans. Amer. Math. Soc., **119**, 1-28.

[313] Pour-El, M. B., Richards, I. (1979). A computable ordinary differential equation which possesses no computable solution. *Ann. Math. Logic*, **17**, 61–90.

[314] Pour-El, M. B., Richards, I. (1981). The wave equation with computable initial data such that its unique solution is not computable. *Adv. in Math.*, **39**, 215–239.

[315] Pour-El, M. B., Richards, I. (1982). Noncomputability in models of physical phenomena. *Int. J. Theor. Phys.*, **21**, 553–555.

[316] Pour-El, M. B., Richards, I. (1989). *Computability in analysis and physics*. Springer-Verlag, Berlin.

[317] Pribram, K. H. (1966). Some dimensions of remembering: steps toward a neuropsychological model of memory. В сб. *Macromolecules and behaviour* (ed. J. Gaito), 165–187. Academic Press, New York.

[318] Pribram, K. H. (1975). Toward a holonomic theory of perception. В сб. *Gestalttheorie in der modern psychologie* (ed. S. Ertel), 161–184. Erich Wengenroth, Kohl.

[319] Pribram, K. H. (1991). *Brain and perception: holonomy and structure in figural processing*. Lawrence Erlbaum Assoc., New Jersey.

[320] Putnam, H. (1960). Minds and machines. В сб. *Dimensions of mind* (ed. S. Hook), New York Symposium. Также в *Minds and machines* (ed. A. R. Anderson), 43–59, Prentice-Hall, 1964; и в *Dimensions of mind: a symposium (Proceedings of the third annual NYU Institute of Philosophy)*, 148–179, NYU Press, 1964.

[321] Ramon y Cajal, S. (1955). *Studies on the cerebral cortex* (пер. на англ.: L. M. Kroft). Lloyd-Luke, London.

[322] Redhead, M. L. G. (1987). *Incompleteness, nonlocality, and realism*. Clarendon Press, Oxford.

[323] Rosenblatt, F. (1962). *Principles of neurodynamics*. Spartan Books, New York.

[324] Roskies, A. (1990). Seeing truth or just seeming true? *Behavioural and Brain Sciences*, **13 (4)**, 682.

[325] Rosser, J. B. (1936). Extensions of some theorems of Gödel and Church. *Jour. Symbolic Logic*, **1**, 87–91.

[326] Rubel, L. A. (1985). The brain as an analog computer. *J. Theoret. Neurobiol.*, **4**, 73–81.

[327] Rubel, L. A. (1988). Some mathematical limitations of the general-purpose analog computer. *Adv. in Appl. Math.*, **9**, 22–34.

[328] Rubel, L. A. (1989). Digital simulation of analog computation and Church's thesis. *Jour. Symb. Logic*, **54 (3)**, 1011–1017.

[329] Rucker, R. (1984). *Infinity and the mind: the science and philosophy of the infinite*. Paladin Books, Granada Publishing Ltd., London. (Первое издание: Harvester Press Ltd., 1982.)

[330] Sacks, O. (1973). *Awakenings*. Duckworth, London.

[331] Sacks, O. (1985). *The man who mistook his wife for a hat*. Duckworth, London.

[332] Sagan, L. (1967). On the origin of mitosing cells. *J. Theor. Biol.*, **14**, 225–274.

[333] Сахаров А. Д. (1967). Квантовые флуктуации вакуума в искривленном пространстве и теория гравитации (Saharov A. D. Vacuum quantum fluctuations in curved space and the theory of gravitation). *Доклады Акад. наук СССР*, **177**, 70–71. Пер. на англ. в *Sov. Phys. Doklady*, **12**, 1040–1041 (1968).

[334] Schrödinger, E. (1935). "Die gegenwartige Situation in der Quantenmechanik". *Naturwissenschaften*, **23**, 807–812, 823–828, 844–849. (Пер. на англ.: J. T. Trimmer (1980) в *Proc. Amer. Phil. Soc.*, **124**, 323–338.) Также в сб. *Quantum theory and measurement* (ed. J. A. Wheeler, W. H. Zurek). Princeton University Press, 1983.



[335] Schrödinger, E. (1935). Probability relations between separated systems. *Proc. Camb. Phil. Soc.*, **31**, 555–563.

[336] Schrödinger, E. (1967). *"What is Life?" and "Mind and Matter"*. Cambridge University Press.

[337] Schroeder, M. (1991). *Fractals, chaos, power laws. Minutes from an infinite paradise*. Freeman, New York.

[338] Scott, A. C. (1973). Information processing in dendritic trees. *Math. Bio. Sci.*, **18**, 153–160.

[339] Scott, A. C. (1977). *Neurophysics*. Wiley Interscience, New York.

[340] Searle, J. R. (1980). Minds, brains and programs. В сб. *The behavioral and brain sciences*. Vol. 3. Cambridge University Press. (Также в сб. *The mind's I* (ed. D. R. Hofstadter, D. C. Dennett). Basic Books, Inc.; Penguin Books Ltd., Harmondsworth, Middlesex, 1981.)

[341] Searle, J. R. (1992). *The rediscovery of the mind*. MIT Press, Cambridge, Massachusetts.

[342] Seymore, J., Norwood, D. (1993). A game for life. *New Scientist*, **139**, No. 1889, 23–26.

[343] Sheng, D., Yang, J., Gong, C., Holz, A. (1988). A new mechanism of high Tc superconductivity. *Phys. Lett.*, **A133**, 193–196.

[344] Sloman, A. (1992). The emperor's real mind: review of Roger Penrose's The Emperor's New Mind. *Artificial Intelligence*, **56**, 355–396.

[345] Smart, J. J. C. (1961). Gödel's theorem, Church's theorem and mechanism. *Synthése*, **13**, 105–110.

[346] Smith, R. J. O., Stephenson, J. (1975). *Computer simulation of continuous systems*. Cambridge University Press.

[347] Smith, S., Watt, R. C., Hameroff, S. R. (1984). Cellular automata in cytoskeletal lattice proteins. *Physica D*, **10**, 168–174.

[348] Smolin, L. (1993). What have we learned from non-pertubative quantum gravity? В сб. *General relativity and gravitation 1992. Proceedings of the thirteenth international conference on GRG, Cordoba, Argentina* (ed. R. J. Gleiser, C. N. Kozameh, O. M. Moreschi). Institute of Physics Publications, Bristol.

[349] Smolin, L. (1994). Time, structure and evolution in cosmology. В сб. *Temponelle scienziae filosofia* (ed. E. Agazzi). Word Scientific, Singapore.

[350] Smorynski, C. (1975). *Handbook of mathematical logic.* North-Holland, Amsterdam.

[351] Smorynski, C, (1983). "Big" news from Archimedes to Friedman. *Notices Amer. Math. Soc.*, **30**, 251–256.

[352] Smullyan, R. (1961). *Theory of Formal Systems*. Princeton University Press.

[353] Smullyan, R. (1992). *Gödel's incompleteness theorem.* Oxford Logic Guide No. 19. Oxford University Press.

[354] Squires, E. J. (1986). *The mystery of the quantum world.* Adam Hilger Ltd., Bristol.

[355] Squires, E. J. (1990). On an alleged proof of the quantum probability law. *Phys. Lett.*, **A145**, 67–68.

[356] Squires, E. J. (1992). Explicit collapse and superluminal signals. *Phys. Lett.*, **A163**, 356–358.

[357] Squires, E. J. (1992). History and many-worlds quantum theory. *Found. Phys. Lett.*, **5**, 279–290.

[358] Stairs, A. (1983). Quantum logic, realism and value-definiteness. *Phil. Sci.*, **50 (4)**, 578–602.

[359] Stapp, H. P. (1979). Whiteheadian approach to quantum theory and the generalized Bell's theorem. *Found. Phys.*, **9**, 1–25.

[360] Stapp, H. P. (1993). *Mind, matter, and quantum mechanics*. Springer-Verlag, Berlin.

[361] Steen, L. A. (ed.) (1978). *Mathematics today: twelve informal essays*. Springer-Verlag, Berlin.



[362] Stoney, G. J. (1881). On the physical units of nature. *Phil. Mag.* (Series 5), **11**, 381.

[363] Stretton, A. O. W., Davis, R. E., Angstadt, J. D., Donmoyer, J. E., Johnson, C. D., Meade, J. A. (1987). Nematode neurobiology using Ascaris as a model system. *J. Cellular Biochem.*, **511A**, 144.

[364] Thorne, K. S. (1994). *Black holes & time warps: Einstein's outrageous legacy*. W. W. Norton and Company, New York.

[365] Torrence, J. (1992). *The concept of nature. The Herbert Spencer lectures*. Clarendon Press, Oxford.

[366] Tsotsos, J. K. (1990). Exactly which emperor is Penrose talking about? *Behavioural and Brain Sciences*, **13 (4)**, 686.

[367] Turing, A. M. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proc. Lond. Math. Soc. (ser. 2)*, **42**, 230–265; исправления в **43**, 544–546.

[368] Turing, A. M. (1939). Systems of logic based on ordinals. *Proc. Lond. Math. Soc.*, **45**, 161–228.

[369] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, **59**, No. 236; также в *The mind's I* (ed. D. R. Hofstadter, D. C. Dennett), Basic Books; Penguin, Harmondsworth, Middlesex, 1981.

[370] Turing, A. M. (1986). Lecture to the London Mathematical Society on 20 February 1947. В сб. *A. M. Turing's ACE report of 1946 and other papers* (ed. B. E. Carpenter, R. W. Doran). The Charles Babbage Institute, vol. 10, MIT Press, Cambridge, Massachusetts.

[371] Tuszńyski, J., Trpisová, B., Sept, D., Satarić, M. V. (1996). Microtubular self-organization and information processing capabilities. В сб. *Toward a science of consciousness: contributions from the 1994 Tucson conference* (ed. S. Hameroff, A. Kaszniak, A. Scott). MIT Press, Cambridge, Massachusetts.

[372] von Neumann, J. (1932). *Mathematische Grundlagen der Quantenmechanik*, Springer-Verlag, Berlin. Пер. на

англ.: *Mathematical foundations of quantum mechanics*. Princeton University Press, 1955.

[373] von Neumann, J., Morgenstern, O. (1944). *Theory of games and economic behaviour*. Princeton University Press.

[374] Waltz, D.L. (1982). Artificial intelligence. *Scientific American*, **247 (4)**, 101–122.

[375] Wang, Hao (1974). *From mathematics to philosophy*. Routledge, London.

[376] Wang, Hao (1987). Reflections on Kurt Gödel. MIT Press, Cambridge, Massachusetts.

[377] Wang, Hao (1993). On physicalism and algorithmism: can machines think? *Philosophia mathematica* (Ser. III), 97–138.

[378] Ward, R.S., Wells, R.O. Jr. (1990). *Twistor geometry and field theory*. Cambridge University Press.

[379] Weber, J. (1960). Detection and generation of gravitational waves. *Phys. Rev.*, **117**, 306.

[380] Weinberg, S. (1977). *The first three minutes: a modern view of the origin of the universe*. Andre Deutsch, London.

[381] Werbos, P. (1989). Bell's theorem; the forgotten loophole and how to exploit it. В сб. *Bell's theorem, quantum theory, and conceptions of the universe* (ed. M. Kafatos). Kluwer, Dordrecht.

[382] Wheeler, J.A. (1957). Assessment of Everett's "relative state" formulation of quantum theory. *Revs. Mod. Phys.*, **29**, 463–465.

[383] Wheeler, J.A. (1975). On the nature of quantum geometrodynamics. *Annals of Phys.*, **2**, 604–614.

[384] Wigner, E.P. (1960). The unreasonable effectiveness of mathematics. *Commun. Pure Appl. Math.*, **13**, 1–14.

[385] Wigner, E.P. (1961). Remarks on the mind-body question. В сб. *The scientist speculates* (ed. I.J. Good). Heinemann, London. (Также в E. Wigner (1967), *Symmetries and reflections*. Indiana University Press, Bloomington; и в *Quantum theory and measurement* (ed. J.A. Wheeler, W.H. Zurek) Princeton University Press, 1983.)



[386] Wilensky, R. (1990). Computability, consciousness and algorithms. *Behavioural and Brain Sciences*, **13 (4)**, 690.

[387] Will, C. (1988). *Was Einstein right? Putting general relativity to the test*. Oxford University Press.

[388] Wolpert, L. (1992). *The unnatural nature of science*. Faber and Faber, London.

[389] Woolley, B. (1992). *Virtual worlds*. Blackwell, Oxford.

[390] Wykes, A. (1969). *Doctor Cardano. Physician extraordinary*. Frederick Muller.

[391] Young, A.M. (1990). *Mathematics, physics and reality*. Robert Briggs Associates, Portland, Oregon.

[392] Zeilinger, A., Gaehler, R., Shull, C.G., Mampe, W. (1988). Single and double slit diffraction of neutrons. *Revs. Mod. Phys.*, **60**, 1067.

[393] Zeilinger, A., Horne, M.A., Greenberger, D.M. (1992). Higher-order quantum entanglement. В сб. *Squeezed states and quantum uncertainty* (ed. D. Han, Y.S. Kirn, W.W. Zachary), NASA Conf. Publ. 3135. NASA, Washington, DC.

[394] Zeilinger, A., Zukowski, M., Horne, M.A., Bernstein, H.J., Greenberger, D.M. (1994). Einstein–Podolsky–Rosen correlations in higher dimensions. В сб. *Fundamental aspects of quantum theory* (ed. J. Anandan, J.L. Safko). World Scientific, Singapore.

[395] Zimba, J. (1993). Finitary proofs of contextuality and nonlocality using Majorana representation of spin-3/2 states, M. Sc. thesis, Oxford.

[396] Zimba, J., Penrose, R. (1993). On Bell non-locality without probabilities: more curious geometry. *Stud. Hist. Phil. Sci.*, **24 (5)**, 697–720.

[397] Zohar, D. (1990). *The quantum self. Human nature and consciousness defined by the New Physics*. William Morrow and Company, Inc., New York.

[398] Zohar, D., Marshall, I. (1994). *The quantum society. Mind, physics and a new social vision*. Bloomsbury, London.

[399] Zurek, W. H. (1991). Decoherence and the transition from quantum to classical. *Physics Today*, **44** (No. 10), 36–44.

[400] Zurek, W. H. (1993). Preferred states, predictability, classicality and the environment-induced decoherence. *Prog. of Theo. Phys.*, **89 (2)**, 281–302.

[401] Zurek, W. H., Habib, S., Paz, J. P. (1993). Coherent states via decoherence. *Phys. Rev. Lett.*, **70 (9)**, 1187–1190.

# ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Интересующие Вас книги нашего издательства можно заказать почтой или электронной почтой:

**subscribe@rcd.ru**

**Внимание:** дешевле и быстрее всего книги можно приобрести через наш Интернет-магазин:

**http://shop.rcd.ru**

Книги также можно приобрести:

1. Москва, ФТИАН, Нахимовский проспект, д. 36/1, к. 307, тел.: 332–48–92 (почтовый адрес: Нахимовский проспект, д. 34)
2. Москва, ИМАШ, ул. Бардина, д. 4, корп. 3, к. 414, тел. 135–54–37
3. МГУ им. Ломоносова (ГЗ, 1 этаж)
4. Магазины:

Москва: «Дом научно-технической книги» (Ленинский пр., 40)
«Московский дом книги» (ул. Новый Арбат, 8)
«Библиоглобус» (м. Лубянка, ул. Мясницкая, 6)
Книжный магазин «ФИЗМАТКНИГА» (г. Долгопрудный, Новый корпус МФТИ, 1 этаж, тел. 409–93–28)

С.-Пб.: «С.-Пб. дом книги» (Невский пр., 28)

**Роджер Пенроуз**

ТЕНИ РАЗУМА: В ПОИСКАХ НАУКИ О СОЗНАНИИ

*Дизайнер М. В. Ботя*
*Технический редактор А. В. Широбоков*
*Корректоры З. Ю. Соболева, М. Г. Пушель*

Подписано в печать 02.08.2005. Формат 84 × $108^{1}/_{32}$.
Печать офсетная. Усл. печ. л. 36,12. Уч. изд. л. 39,94.
Гарнитура Литературная. Бумага офсетная №1.
Тираж 1500 экз. Заказ № 3963.

АНО «Институт компьютерных исследований»
426034, г. Ижевск, ул. Университетская, 1.
http://rcd.ru E-mail: borisov@rcd.ru

Отпечатано в полном соответствии с качеством
предоставленных диапозитивов в ОАО «Дом печати — ВЯТКА»
610033, г. Киров, ул. Московская, 122