

VEcordia

Извлечение R-PENRS2

Открыто: 2010.08.22 16:32
Закрито: 2010.09.04 23:04
Версия: 2016.12.10 15:48

ISBN 9984-9395-5-3

Дневник «VECORDIA»

© Valdis Egle, 2016

ISBN 5-93972-457-4

Роджер Пенроуз. «Тени Разума», том II

© Роджер Пенроуз, 1994



Shadows of the Mind

Роджер Пенроуз

ТЕНИ РАЗУМА

Том 2
(главы 3 – 4)

С комментариями Валдиса Эгле

Impositum

Grīziņkalns 2016

Talis hominis fuit oratio,
qualis vita

Предисловие к Третьей главе Пенроуза

2010.08.24 16:30 вторник

Третья глава, к которой я так спешил и от которой так много ожидал (см. конец книги {PENRS1}), надо признать, меня несколько разочаровала. Я ожидал увидеть там примеры из математической практики, в которых (по Пенроузу) наблюдается «невывислимость математического мышления», и я тогда мог бы эти примеры разобрать в свете Веданской теории.¹

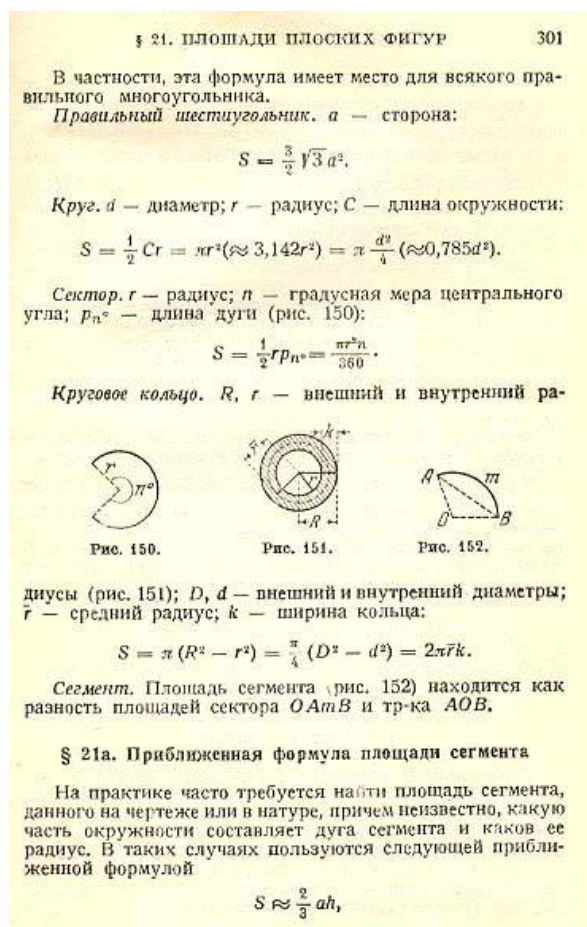


Рис. VE1. М.Я. Выгодский. Справочник по элементарной математике. Издание четырнадцатое. Государственное издательство физико-математической литературы, Москва, 1962



Рис. VE2. И.Н. Бронштейн и К.А. Семендяев. Справочник по математике для инженеров и учащихся ВТУЗов. Издание десятое, стереотипное. Издательство «Наука», Москва, 1964

Но вместо этого в Третьей главе Пенроуза оказались почти сплошные разглагольствования о «формальных системах» Q , Q^* , $Q(M)$, $Q_m(M)$, $Q'_m(M)$, $Q'_m(M,c)$ и т.д.

Я вынужден повторить (см. конец книги {CANTO2}), что «формальные системы» вообще не являются математикой и не принадлежат ей. Чтобы напомнить читателю, что такое математика, я взял четыре тома из своей библиотеки, открыл их на случайных страницах, засканировал эти страницы и представил читателю на рисунках VE1, VE2, VE3 и VE4.

¹ P.S. 2010.09.02: Хороший пример — теорему Гудстейна — Пенроуз дал во «Вступлении» 1998 года к повторному изданию своей Первой книги (см. {PENRO1}).

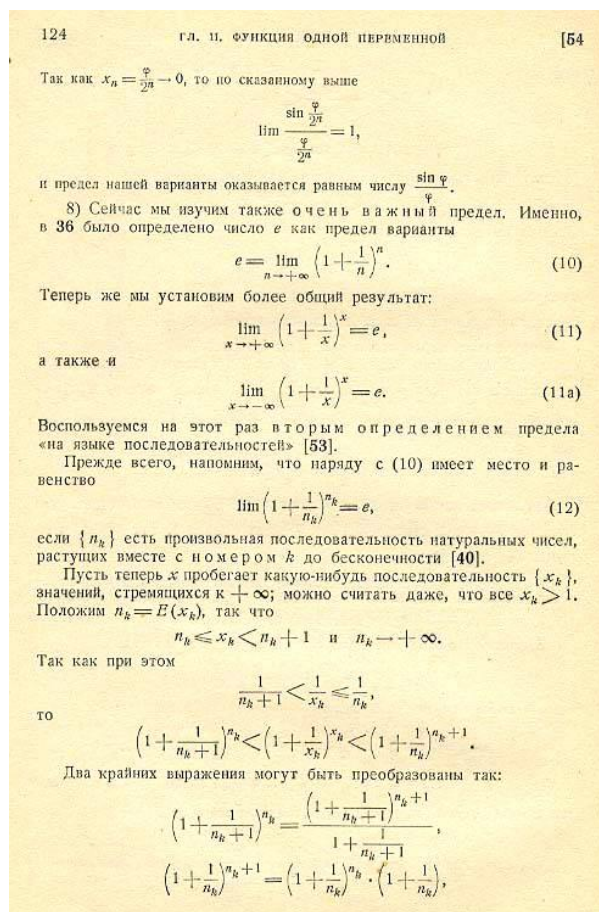


Рис. VE3. Г.М. Фихтенгольц. Курс дифференциального и интегрального исчисления. Том I. Издание четвертое, исправленное. Государственное издательство физико-математической литературы, Москва, Ленинград, 1958

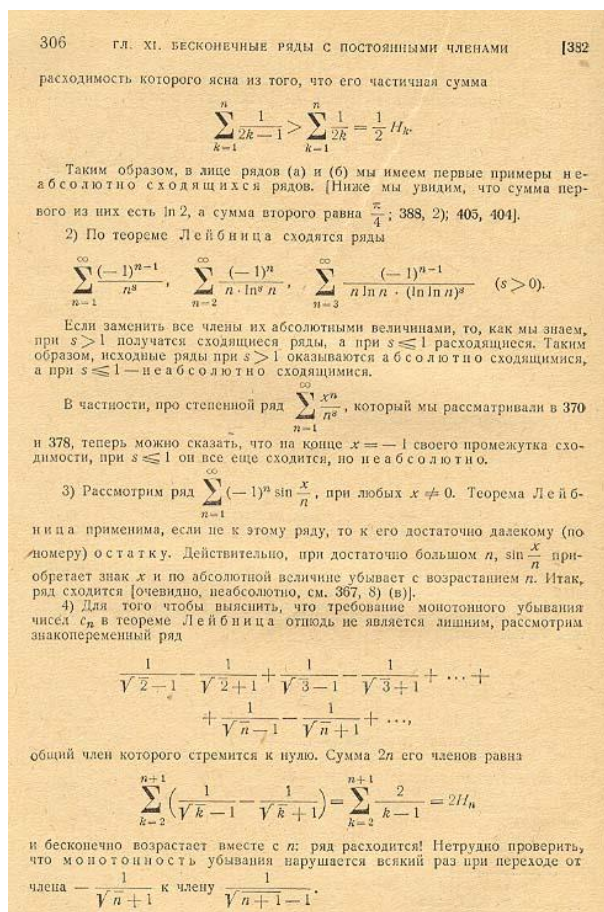


Рис. VE4. Г.М. Фихтенгольц. Курс дифференциального и интегрального исчисления. Том II. Издание четвертое, исправленное и дополненное. Государственное издательство физико-математической литературы, Москва, Ленинград, 1959

Вот ЭТО называется математикой (если кто забыл).

Это дает мне (и всем другим людям) возможность вычислить и найти разные вещи, начиная от простой площади круга или объема цилиндра до самых хитроумных интегралов и решений всяких там дифференциальных уравнений.

Это изумительно красивая наука, которой я восхищался с детства.

Но что же такое «формальные системы»? Возможность что вычислить или найти дают ОНИ мне?

Ничего!

Так зачем они нужны? Зачем они на свет появились?

Оказывается, та красивая наука, которая называется математикой, недостаточно хорошо обоснована.

И вот, эти «формальные системы» занимаются ОБОСНОВАНИЕМ математики.

Что ж — в принципе это неплохое и в конце концов, может быть, даже и полезное занятие. Если только они знали бы, что такое математика из себя представляет, откуда возникает, ЧТО действительно лежит в ее основании.

Вся беда заключается в том, что разработчики этих «формальных систем» НЕ ЗНАЛИ, что такое математика, откуда она появилась и что из себя представляет. Поэтому их попытки «обосновать» математику пошли в фундаментально неправильном направлении, глобально «не в ту степь», очень скоро они завязли в безнадежных болотах, запутались в противоречиях, и в общем и в целом ВСЁ ИХ НАПРАВЛЕНИЕ деятельности не имеет никакой ценности.

Это тупиковый путь, бесплодная ветвь, которую давно надо отрезать от дерева математики и выбросить в кучу мусора.

А разбирать надо подлинные основания математики, которые показывает Веданская теория. Подлинные основания математики — это некоторый набор мозговых программ, и

изучение математики есть изучение этих программ. Так и надо ставить вопрос об основаниях математики, а не о каких-то там аксиомах и формализации рассуждений о них.

Эту глобальную ориентацию нужно иметь в виду, разбирая текст Пенроуза.

Он не решает математические проблемы в этой Третьей главе; он бродит по непроходимым болотам глобально неверного пути, по которому заблудилась математическая мысль в конце 19-го столетия.

Трудно вообще комментировать, когда разбираемый текст весь лежит на фундаментально неверной почве. Этой неправильной основе вообще-то надо просто противопоставить правильную основу. Но это не задача настоящей книги. Задача настоящей книги – получить, зафиксировать и хотя бы частично осветить текст Пенроуза. ЕГО воззрения, а не мои, всё же являются главным предметом этого тома.

Поэтому далее даю ЕГО текст и комментирую уж настолько, насколько это возможно.

Роджер Пенроуз. «Тени разума»

(Продолжение; начало в книге {PENRS1})

Глава 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 $\mathcal{G}$), суть которого заключается в том, что математическое понимание не может являться результатом применения какого-либо осмысленно осознаваемого и полностью достоверного алгоритма² (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 $\mathcal{G}$, 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.³ 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»⁴; столь же непреложными, на их взгляд,

² В.Э.: В связи с пенроузовским утверждением $\mathcal{G}$ у меня теперь создалось двойственное положение: с одной стороны, я признаю его истинность в той интерпретации, что «математическое понимание не может являться результатом применения какого-либо алгоритма» формалистских процедур (которые сам Пенроуз на самом деле и имеет в виду в своем утверждении $\mathcal{G}$); с другой стороны – это утверждение не имеет никакого отношения к «доказательству», данному им в §2.5 {PENRS1}.

³ В.Э.: Пенроуз пытается «охватить все возможные случаи», однако это ему никак не удается, потому что это не позволяет ему та математическая парадигма, в рамках которой он рассуждает. Он просто не видит тот случай, который имеет место в действительности: «убежденность математика в истинности своих выводов» появляется не от применения какого-то «неизвестного и неосознаваемого алгоритма», а от изучения алгоритма (мозговой программы), о существовании которого традиционная математика (и Пенроуз), правда, в настоящее время не догадываются, но который вполне постижим после того, как его предьявила Веданская теория.

⁴ В.Э.: ...и представляющие собой знания об алгоритмах мозговых программ.

являются и все промежуточные умозаключения, составляющие упомянутую последовательность. Какой бы длинной, запутанной или даже концептуально неочевидной ни была цепь умозаключений, само рассуждение⁵ 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.⁶ 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.⁷

Весьма поучительным представляется рассмотреть точки зрения двух выдающихся мыслителей от математики, которым мы, собственно говоря, и обязаны идеями, приведшими нас к утверждению $\mathcal{G}$. 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 $\mathcal{G}$. 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 напроць 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,⁸ не считал, что утверждение $\mathcal{G}$ 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».

⁵ В.Э.: Пенроузовская парадигма акцентирует рассуждения (берущие начало, конечно же, из аксиом!). А следовало бы акцентировать знания о некоторых объектах (мозговых программах), на самом деле являющиеся стартовыми для всех этих рассуждений.

⁶ В.Э.: Существует не неизвестная математикам процедура (рассуждений), а неизвестный математикам объект (изучения).

⁷ В.Э.: Да – такой фактор существует. И он представляет собой именно «алгоритмически описываемые процедуры»; но весь «фокус» в том, что это не те процедуры рассуждений (которые Пенроуз ищет и ищет и ищет...), а совсем другие процедуры, которые являются предметом математических знаний и которые математика изучает «со стороны».

⁸ 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.

Надо сказать, что вышеприведенное допущение ни в коей мере не противоречит $\mathcal{G}$ (и я ничуть не сомневаюсь, что Гёделю был хорошо известен тот недвусмысленный вывод, какой в моей формулировке получил обозначение $\mathcal{G}$). 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 $\mathcal{D}$), 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 лице утверждения $\mathcal{G}$. 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 ошибаться⁹:

«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».¹⁰

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

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,¹¹ 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 $\mathcal{G}$, 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, имея они шанс хоть сколько-нибудь всерьез предположить, что физический процесс может иногда оказаться в основе

⁹ В.Э.: Роботам ошибаться нет необходимости «позволять»; как только компьютер станет самопрограммирующимся, ошибок будет предостаточно, а проблемой станет нахождение правильного решения.

¹⁰ См. [198], с. 361. Цитата взята из лекции Тьюринга, прочитанной в 1947 году перед Лондонским математическим обществом и приводится по изданию [370].

¹¹ В.Э.: За счет самопрограммирования; но это самопрограммирование порождает и ошибки.

своей невычислимым – в соответствии с точкой зрения $\mathcal{C}$, ради продвижения которой и была написана эта книга.

В последующих разделах (особенно, в §§3.2–3.22) я представляю вашему вниманию несколько детальных обоснований (некоторые из них довольно сложны, запутаны или специальные), целью которых является демонстрация неспособности вычислительных моделей $\mathcal{A}$ и $\mathcal{B}$ 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) всё же начать чтение, а затем, когда уж совсем надоест, переходить сразу к итоговому воображаемому диалогу (§3.23). Если у вас затем снова 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 $\mathcal{C}$ для того, чтобы математическое понимание могло оказаться результатом выполнения некоего алгоритма, этот алгоритм должен быть необоснованным или непознаваемым,¹² 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 компьютере¹³ 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 единственный тип алгоритмических процедур, а именно – нисходящие. Такую алгоритмическую процедуру можно относить как к отдельному математику, так и к математическому сообществу в

¹² В.Э.: «Математическое понимание», конечно же, является «результатом выполнения некоего алгоритма». Ситуация такова, что 1) имеются (мозговые) программы (и встроенные в них алгоритмы) группы А, представляющие собой предмет математики; и 2) имеются (мозговые) программы (и встроенные в них алгоритмы) группы В, которые изучают эту группу А «со стороны» и в результате приобретают математические знания и математическое понимание. (Примером программы группы А может послужить «гудстейнатор», с подачи Пенроуза разобранный мною в {PENRO1}, а программами группы В – упомянутые там же в пунктах 1–7 программы {PENRO1}). Вариант «непознаваемые» относительно этой группы В можно отбросить сразу как заведомую ерунду; вариант «необоснованный» требует уточнения. Конечно, алгоритмы этой группы В (как и группы А) не принадлежат к используемым и рассматриваемым математическими формалистами процедурам. В этом смысле они действительно «необоснованы». Но они могут быть изучены (без помощи математического формализма) и знания о них обоснованы (без кавычек), как и любые другие научные знания.

¹³ В.Э.: Рассуждения Пенроуза верны лишь в том случае, если под «алгоритмом» понимать только процедуры математических формалистов для их рассуждений (как это Пенроуз на самом деле и понимает); если же под алгоритмом понимать «любой набор операций, который можно, в принципе, смоделировать (на самом деле: не смоделировать, а выполнить) на универсальном компьютере», то все рассуждения Пенроуза превращаются в чушь. Ну не относится всё то, что он говорит, к алгоритмам тех групп А и В, о которых я говорил в предыдущей сноске! (Какие в отношении их «варианты I, II и III»? – такие варианты вообще, в принципе, могут существовать только в рамках формалистской парадигмы, где знания и «понимание» получают лишь путем «рассуждений», исходящих из аксиом и т.д. В этой парадигме Пенроуз и сидит всё время, и не может из нее выскочить).

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, повторяюсь, 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 Π_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 $\mathcal{G}^{**}$ 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 F .¹⁴ Иными словами, получается, что математическое понимание – или хотя бы понимание математики каким-то отдельным математиком – эквивалентно выводимости в рамках некоторой формальной системы F . Если мы хотим сохранить хоть какую-то надежду удовлетворить выводу $\mathcal{G}$, к которому нас столь неожиданно привели изложенные в предыдущей главе соображения, то придется предположить, что система F является необоснованной. Однако, как это ни странно, необоснованность в данном случае ситуацию ничуть не меняет, поскольку, в соответствии с I, известная формальная система 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 F . (Согласитесь, крайне неразумно выглядит точка зрения, в соответствии с которой математик позволяет себе не верить в самые фундаментальные

¹⁴ В.Э.: Вот, вот – алгоритм эквивалентен формальной системе!.. Это вообще типичное и фундаментальное заблуждение всех формалистов. Профессор Подниекс в «Канториане» четверть века назад утверждал мне то же самое {[CANTO2.1787](#)}: «*всякую программу можно преобразовать в эквивалентный ей список аксиом и правил вывода*». В ответ {[CANTO2.1957](#)} я предложил ему формализовать мою программу вычисления биоритмов. Само собой разумеется, что Подниекс не только не сделал этого; он даже не попытался начать, даже не попытался указать способ, каким это (хотя бы в принципе) можно было бы сделать; вместо этого он больше никогда не касался этого вопроса, избегал всякого упоминания о нем... Так что чушь это! (В каких-то «чисто, сугубо, абсолютно теоретических» заоблачных высях, может быть (?), и вправду можно из программы (и, соответственно, из ее алгоритма) сделать какую-нибудь невообразимо сложную, запутанную и неуклюжую формальную систему, с которой – в отличие от работающей программы и работающего алгоритма – никто не сможет ничего предпринять. В таком случае вступает в силу несостоятельность «доказательства» Пенроуза из [§2.5](#), которая уничтожает все его рассуждения. Но так как практически алгоритм не эквивалентен формальной системе, то с практической точки зрения нам здесь лучше разделять оба эти понятия и считать, что существуют (1) алгоритмы, которые вытекают из (реальных) формальных систем, и (2) алгоритмы, которые (во всяком случае на практике) не эквивалентны никаким формальным системам. Названные мною выше алгоритмы группы В (которыми в действительности достигается «математическое понимание») тогда принадлежат к группе (2), а Пенроуз говорит о группе (1) (считая ее единственной существующей). Такой моделью я в основном и руководствуюсь).

положения собственной заведомо непроверяемой системы взглядов.) Независимо от того, является ли система F действительно обоснованной, вера в ее обоснованность уже содержит в себе веру в то, что утверждение $G(F)$ (или, как вариант, $\Omega(F)$, см. §2.8) истинно. Однако, поскольку теперь мы полагаем (исходя из веры в справедливость теоремы Гёделя), что истинность утверждения $G(F)$ в рамках системы F недоказуема, это противоречит предположению о том, что система 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 Π_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 F обоснованной или нет. Если система F сама по себе познаваема, то следует рассмотреть возможность II, суть которой заключается в том, что система F всё же может составлять основу математического понимания, однако узнать об этой ее роли мы не в состоянии. Остается в силе и возможность III: сама система 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 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 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 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 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 F , будучи, возможно, крайне замысловатой, всё 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 F действительно целиком и полностью охватывает всю совокупность наших твердых математических убеждений и интуитивных озарений (по крайней мере в том, что касается Π_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.¹⁵ Мы еще поговорим об этом в конце данного раздела и – более подробно – в §§3.15 и 3.29.

¹⁵ В.Э.: Чтобы создать робота-математика, надо в него встроить те группы A и B программ, о которых я говорил в предыдущих сносках, а не эту белиберду про «системы F » и « Π_1 -высказывания». А встроить те группы программ A и B не представляет никаких теоретических трудностей.

Дабы подчеркнуть тот факт, что существование подобной системы 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 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 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. Ибо если мы не допускаем, что процедура F целиком и полностью обоснована, то это означает, что мы отвергаем какие-то ее следствия. А ее следствиями являются как раз те математические положения (или хотя бы только Π_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 Π_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 μ -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 Π_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 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 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 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 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 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 F неизбежно имеются такие правила, неоспоримая обоснованность которых не может быть достоверно установлена ни одним математиком – причем мы всё еще полагаем, что число аксиом в системе F конечно.

В чем же причина? Перенесемся в воображении в то самое время, когда уже однозначно установлена неопровержимая справедливость всех аксиом формальной системы F. Перед нами открывается замечательная возможность без помех рассмотреть всю систему F целиком. Попробуем допустить, что все правила действия системы F можно также считать справедливыми безо всяких оговорок. Хотя предполагается, что мы еще не можем знать наверняка, что система F действительно включает в себя всю математику, которая в принципе доступна человеческому пониманию и интуиции, мы должны к настоящему моменту, по меньшей мере, уже убедиться в том, что система F является 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 F непротиворечива. Не забываем, разумеется, и о том, что, в силу этой непротиворечивости, утверждение $G(F)$ также должно быть истинным – более того, неопровержимо истинным! Однако, поскольку предполагается, что система фактически (хотя нам об этом неизвестно) включает в себя всю совокупность того, что безоговорочно доступно нашему пониманию, утверждение $G(F)$ должно на деле представлять собой теорему системы F. Согласно теореме Гёделя, такое, вообще говоря, возможно только в том случае, если формальная система F противоречива. Если же система 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 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 F с конечным набором аксиом безоговорочно истинны, нам остается лишь признать, что противоречие действительно имеет место. Следовательно, среди правил действия системы 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 F задается конечным набором аксиом. В качестве возможного альтернативного решения можно предположить, что количество аксиом в системе F бесконечно. Относительно этой возможности необходимо сделать некоторые комментарии. Для того чтобы систему 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 ZF) – включает в себя бесконечное количество аксиом, выражаемых посредством структур, называемых «схемами аксиом». Путем соответствующего переформулирования систему ZF можно выразить таким образом, что количество действительных аксиом станет конечным)¹⁶. Более того, действуя определенным образом, такое можно проделать с любой схемой аксиом, являющейся «формальной» в требуемом нами вычислительном смысле.¹⁷

Может создаться впечатление, что вышеприведенное рассуждение (целью которого является исключение из списка возможных вариантов случая II) применимо к любой (обоснованной) системе 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 столь самоочевидно обоснованными. Так, представляя себе, в соответствии с вышеизложенными соображениями, времена, когда нам станут известны все аксиомы и правила действия системы 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 неопровержимо установления обоснованности правил действия такой системы 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 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 F (на основании которой человек постигает истинность Π_1 -высказываний), целиком и полностью понимаемая математиками, обладающая конечным набором аксиом, справедливость которых не вызывает никаких сомнений, и конечной системой правил действия $\mathcal{R}$, которая, впрочем, содержит по крайней мере одну операцию, полагаемую фундаментально сомнительной. Каждая отдельно взятая теорема системы 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 $\mathcal{R}$. Кроме того, хотя математик и может (в принципе) установить истинность каждой из упомянутых теорем в отдельности, единообразно процедуры для этого не существует. Можно ограничить область рассмотрения теми теоремами системы F, которые представляют собой Π_1 -высказывания. Применяя сомнительную систему правил $\mathcal{R}$, мы можем вычислительным способом сгенерировать перечень тех Π_1 -высказываний, справедливость которых может быть однозначно установлена математиками. В конечном счете, человек, воспользовавшись пониманием и интуицией, оказывается способен установить справедливость каждого из этих Π_1 -высказываний в отдельности. Однако в каждом конкретном случае для такого установления применяются методы рассуждений, существенно отличающиеся от правила с помощью которого было получено данное Π_1 -высказывание. Раз за разом нам приходится добавлять в систему всё новые, всё более изощренные плоды человеческого разума – с тем, чтобы можно было неопровержимо доказать истинность каждого последующего Π_1 -высказывания. Словно по волшебству, истинными оказываются все Π_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 всё более глубоких уровнях.¹⁸ Более того,

¹⁶ Упомянутая процедура заключается во вложении системы ZF в систему Гёделя–Бернайса; см. [56], глава 2.

¹⁷ 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.

¹⁸ В.Э.: Весь разговор Пенроуза строится на фундаментальном предположении, что математика представляет собой совокупность «высказываний» (аксиом, теорем), получаемых при помощи «правил вывода», и тогда исключительную важность получают все эти оценки «обоснованности», «непротиворечивости» и т.д. Но это фундаментальное предположение ошибочно. Математика (та математика, четыре

любое Π_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 $\mathcal{K}$. Наконец, существует еще и особое истинное Π_1 -высказывание $G(F)$, которое явным образом выводится из знания формальной системы F , однако истинность которого не может быть неопровержимо установлена ни одним математиком. В лучшем случае, математик сможет понять, что истинность $G(F)$ непосредственно обусловлена обоснованностью сомнительной системы правил действия $\mathcal{K}$, которая, по всей видимости, обладает некоей чудесной способностью определять, истинность каких именно Π_1 -высказываний может быть неопровержимо установлена человеком.

Могу себе представить, что кому-то всё это, возможно, покажется не совсем бессмысленным. Ко многим своим выводам математики приходят на основании предпосылок, которые можно назвать «эвристическими принципами»¹⁹ – 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 $\mathcal{K}$. В сущности, такие принципы способны лишь углубить наше сознательное понимание причин, в соответствии с которыми оказывается истинным тот или иной математический вывод.²⁰ Впоследствии, в результате более серьезной разработки 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 Π_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 F), описанного в соответствии с возможностью II. Более того, никто никогда и не предлагал эвристического принципа, позволившего бы сгенерировать в точности все Π_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

странички из которой я отобразил в Предисловии) не является совокупностью «высказываний» от «правил вывода». Она является системой знаний об определенном объекте (группа программ A в моих предыдущих сносках), и как таковая не отличается слишком существенно от систем знаний в других областях наук (например, в биологии или в программировании). Мы же не строим черт знает какие пирамиды «высказываний» и «правил вывода», чтобы отобразить свои знания о системе, например, WINDOWS, и не заботимся до иступления о «непротиворечивости» и «обоснованности» этих знаний. Конечно, наши знания о WINDOWS должны быть и обоснованными и непротиворечивыми, но нет того страшного зуда, который терзает всех формалистов. И если отбросить то фундаментальное предположение, которым Пенроуз здесь руководствуется, и увидеть, наконец, вещи такими, какие они есть на самом деле, то также и из математики исчезнет этот постоянный зуд «обоснованности» и «непротиворечивости», и она займется – как в прежние времена – просто изучением математических истин – подобно всем остальным наукам.

¹⁹ В.Э.: ...но которые на самом деле являются изучением программ группы A (в терминах моих предыдущих сносков).

²⁰ Эвристический принцип такого рода может принять форму гипотезы – в качестве примера укажем весьма значительную гипотезу Таиямы (обобщенную позднее в так называемую «философскую теорию Лэнгленда»), в виде следствия из которой можно представить самое, пожалуй, знаменитое из Π_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 « $\mathcal{K}$ », но рассуждение, доказывающее (в соответствующем случае) саму гипотезу Таиямы!

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 того, принадлежит он к лагерю $\mathcal{A}$ или $\mathcal{B}$) является в высшей степени безрассудным возлагать какие бы то ни было надежды на отыскание такой алгоритмической процедуры²¹ (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.²²

Можно ли допустить, что подобный алгоритм F всё 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.²³

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 F) может оказаться необоснованной (случай I, как мы помним, такой лазейки не допускал). Может ли быть так, что математическое понимание человека представляет собой эквивалент некоего познаваемого алгоритма, который в основе своей ошибочен? Рассмотрим эту возможность подробнее.

§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 F . Как же мы тогда можем быть уверены в том, 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 F и в самом деле является необоснованной. Здесь же мы допускаем, что подобная роль системы F принципиально непознаваема, вследствие чего нам придется повторно рассмотреть вариант с возможной необоснованностью 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

²¹ 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.

²² В.Э.: Нет никаких теоретических препятствий к тому, чтобы воплотить в компьютер группы программ A и B , упомянутые в моих предыдущих сносках и порождающие у людей математические знания и математическое понимание. Однако принадлежат ли они к пенроузовской «процедуре F »? С одной стороны – нет, ибо «процедура F » – это «машина для доказательства теорем», а программы групп A и B вообще-то не это. С другой стороны – A и B к «процедуре F » принадлежат, ибо для Пенроуза F – это любая «вычислительная процедура», порождающая у людей математические знания и математическое понимание. Вся постановка проблемы у Пенроуза некорректна! (И вытекает она из фундаментального – и неверного – предположения о том, что математика – это совокупность «высказываний»... и т.д.).

²³ В.Э.: Бедные мы! ☺

принципе, оказаться уверенностью 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, они даже сошлутся 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 нет? Всё 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 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.²⁴ 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-

²⁴ В.Э.: Самым фундаментально неверным, конечно, было то, что Фреге вообще взялся за создание формальной системы. Это (как и одобрение такого пути Гильбертом и другими) было катастрофой для математики. Она ушла по неправильному пути в болота, где ее ожидали непроходимые топи и сплошные прорывы; она перестала быть точной и достоверной наукой, а жизни сотен, тысяч людей были истрачены впустую на бесплодную, никому не нужную работу, не приносящую людям никакой пользы... Но винить Фреге, Гильберта и других тоже трудно: настоящий, правильный путь в то время и в самом деле не просматривался – до изобретения компьютеров оставалось еще полвека. Ну как они могли догадаться, что надо не строить формальные системы, а изучать мозговые программы!

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 осознаем лишь тогда, когда такой парадокс наконец себя проявит. Короче говоря, наша математическая интуиция не зиждется на каких-то непреходящих в веках установлениях, но 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 всё 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), пришедшую 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.²⁵ (Следует, впрочем, отметить, что сам Кантор знал о парадоксах, подобных парадоксу Рассела, – задолго до того, как сам Рассел обнаружил тот, что был назван его именем,²⁶ – 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

²⁵ В.Э.: Что за ерунда?! Кантор свои идеи о бесконечных множествах выдвинул в 1870-х годах, что отнюдь не начало XIX века; общепринятыми они стали, можно сказать, с начала XX века, но это не время их выдвигания Кантором.

²⁶ См. [181], с. 74.

качестве основы для своего математического понимания и убеждений какую-либо необоснованную формальную систему 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 F, вернемся ненадолго к другим аспектам человеческой «неточности», о которых мы говорили выше (см. комментарии к возражениям [Q12](#) и [Q13](#)). Прежде всего повторюсь, 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 неожиданному, 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 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 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–Тьюринга²⁷; на этом, собственно, основании мы и не рассматриваем всерьез возможность I, а возможность II полагаем крайне невероятной. Как уже отмечалось выше (в комментариях 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(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

²⁷ В.Э.: То есть – ту «муру», которая была дана в [§2.5](#)?! Это обязательно должен понимать математик?! (Да любой трезвомыслящий человек должен понимать, что на «диагональном методе» никакие доказательства строить нельзя!)

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 Π_1 -высказывания доступна математическому пониманию человека, мы, по сути, утверждали, что данное Π_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 Π_1 -высказываний. Π_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 Π_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 Π_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 определяемый временной промежуток (в масштабе времени Планка равный приблизительно $0,5 \times 10^{-43}$ с, см. §6.11), то и в этом случае за всё время существования Вселенной, начиная от «большого взрыва» и до настоящего момента, нам не удалось бы

добраться ни до какого 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 «непознаваемыми» 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 в качестве алгоритма, которым, как 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.²⁸ За пределами человеческих возможностей оказывается 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, $\mathcal{A}$ так и $\mathcal{B}$)? Адепты полностью автоматизированных ИИ-систем (т.е. сторонники непременно $\mathcal{A}$, а также, возможно, кто-то из лагеря $\mathcal{B}$) 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 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,²⁹ возможно, несколько более примитивных, нежели ожидаемые нами роботы-математики; кроме того, в процессе развития роботов будет, возможно, принимать участие и

²⁸ 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.

²⁹ См. [267], [268].

некое подобие дарвиновской эволюции, в результате чего от поколения к поколению роботы будут становиться всё более совершенными. Разумеется, не обойдется 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 $\mathcal{A}/\mathcal{D}$, 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 объяснить³⁰ с позиций той науки, которая достигла столь значительных успехов в описании мира неодушевленных предметов. Разумеется, мы по-прежнему должны 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 $\mathcal{D}$. 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 всё 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 $\mathcal{D}$. Надеюсь, что те читатели, которых можно отнести к приверженцам той или иной формы $\mathcal{D}$, всё же потерпят меня еще некоторое время, а я пока попробую выяснить, к чему нас может привести в данном случае научный подход.

³⁰ В.Э.: Но ведь можно же!

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 образования 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 собраны наши тела, равно как 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 один или их много?

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 по-разному. 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, с 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)³¹. Однако часто бывает так, что всеми этими способами воспринимается одна и та же математическая истина. С алгоритмической точки зрения первое впечатление таково, что 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 ζ ***). 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, обуславливающих осознание человеком истинности какого-либо математического утверждения, эти процессы должны весьма и весьма различаться от индивидуума к индивидууму,

³¹ См., напр., [102] (и НПК, глава 9).

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.³²

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), обуславливающий 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 часто 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.³³ Если уж мы твердо

³² В.Э.: Совершенно верно! Математические истины – это знания о (мозговых) программах группы А (например, таких как «гудстейнатор» из *Вступления* книги {PENRO1}). Однако часто эти программы бывают в различных близких вариантах (например, натуральные числа определяются как программой N, классифицирующей множества по количеству элементов и помещающей в соответствующий таксон множество из трех голубей; – так и программой R, классифицирующей соотношения множеств, и помещающей в соответствующий таксон отношение трех голубей к одному голубю – т.е. пару двух множеств: из трех и из одного голубя). Более того, многие математические соотношения проявляются в продуктах (мозговых) программ совершенно различной природы (оперирующих с различными объектами и выполняющих разную работу). В таком случае для «визуализации» данного математического соотношения можно использовать и ту, и другую, и третью из этих программ (и разные люди могут предпочитать разные программы для своих «визуализаций»). Так и появляются «геометрическая интерпретация», «алгебраическая интерпретация» и т.д.

³³ В.Э.: Черт возьми! Как можно столько блуждать в трех соснах!

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 сколь угодно заметных практических преимуществ.³⁴ 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 всё 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.³⁵ 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, возможно, как постройка убежищ, изготовление одежды, изобретение ловушки для мамонтов или, несколько позднее, одомашнивание животных и выращивание урожая. (См. рис. 3.1.)

Разумно было бы предположить, что упомянутые преимущества, которыми, очевидно, всё 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 отнюдь не уникален в своей способности понимать. Такой же способностью обладали, возможно, и многие другие

³⁴ В.Э.: Это не так. Для конкретности возьмем, например, натуральные числа, порождаемые программой N. Эта программа (см. {PENRS1}) классифицирует множества (т.е. объекты внешнего мира) по количеству элементов – и как таковая принадлежит к целому обширному классу программ классификации множеств вообще по различным всевозможным признакам: на съедобные и несъедобные, на опасные и безвредные и т.д. Польза от совершенствования программ классификации объектов внешнего мира для индивида очевидна – потому эти программы и шлифовались естественным отбором, пока среди алгоритмов различных классификаций не появилась и классификация по количеству элементов, положившая начало математике. Так что польза от этого направления умственной деятельности была уже и до появления программы N (т.е. натуральных чисел). Но после ее появления опять же была очевидная польза: те племена, которые умели считать, могли организовать более высокоразвитое общество – со сбором налогов (самое первое применение арифметики), с централизованным войском и т.д. – и побеждать другие племена в жестокой борьбе за передел Планеты.

³⁵ В.Э.: Это не так. В ДНК надо было кодировать не конкретно программу N, а способность мозга создавать вообще различные классификации множеств – гибкость мозга и его вычислительную мощь.

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.

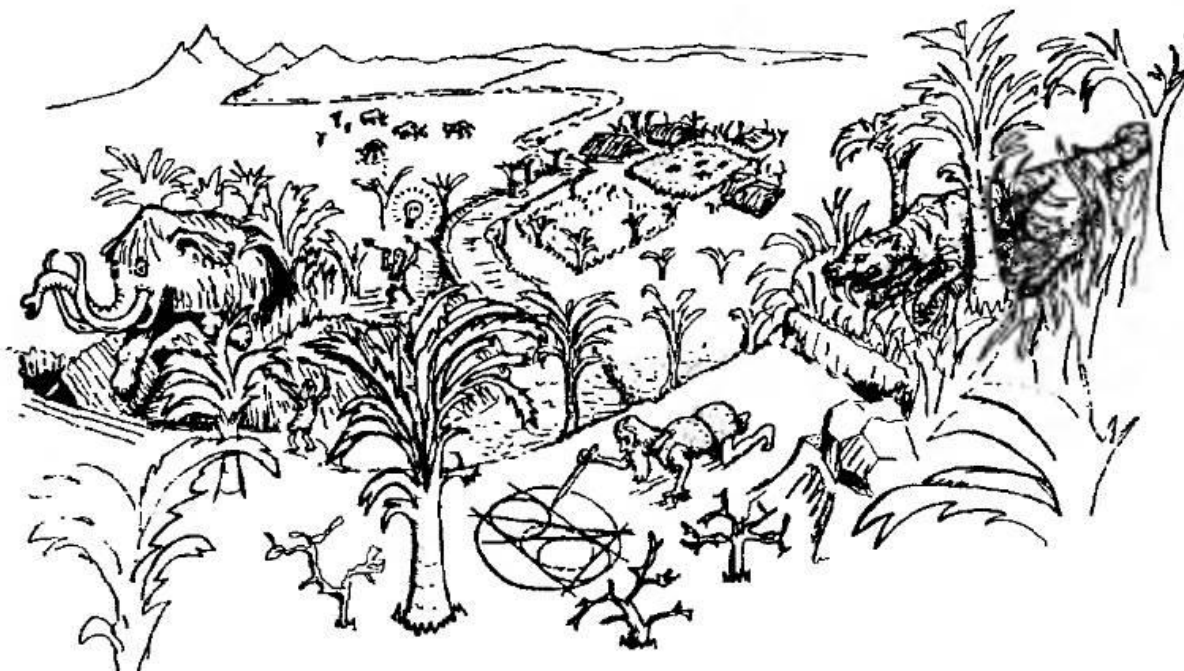


Рис. 3.1. Вряд ли специфическая способность составлять сложные математические суждения могла дать нашим далеким предкам какие бы то ни было преимущества в борьбе за существование, а вот общая способность к пониманию им наверняка не помешала бы.

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 наших далеких предков он уже включал в область своей потенциальной применимости правила всех формальных систем,³⁶ рассматриваемых сегодня математиками как безоговорочно непротиворечивые (или неопровержимо обоснованные, если речь идет о Π_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, которые могут быть получены 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 рассуждений³⁷ – 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

³⁶ В.Э.: Ну вот – Пенроуз сам и опровергает формализм! Конечно, не «формальные системы» (существовавшие якобы в каком-то скрытом виде) были полезны для естественного отбора, а – мозговые программы.

³⁷ В.Э.: Боже мой! – какой вздор!

различий, отделению справедливых аргументов от ничем не обоснованных во всех тех, тогда еще не открытых, областях математики, которые сегодня оккупируют³⁸ страницы специальных научных журналов. Все вышеперечисленные способности должны были оказаться каким-то образом закодированы внутри этого самого – гипотетического, непознаваемого или, если хотите, 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»³⁹ имеет неалгоритмическую природу,⁴⁰ как ситуация в корне меняется. Теперь уже нет необходимости приписывать этой способности какую-то невероятную 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,⁴¹ что в принципиальную невозможность корректного моделирования понимания посредством какой бы то ни было вычислительной процедуры верится с трудом. И всё 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,⁴² 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 всё, 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 обучаться. Такая система может состоять, в основе своей, из «восходящих» компонентов, соединенных по мере необходимости с какими-либо «нисходящими» процедурами (см. §1.5)⁴³.

³⁸ В.Э.: просто занимают. (Здесь поработал «ложный друг переводчика»...)

³⁹ В.Э.: ...т.е. строить гибкие и мощные мозговые программы...

⁴⁰ В.Э.: Эта пенроузовская «неалгоритмическая природа» есть на самом деле мозговые программы, работающие «не по правилам» математических формалистов.

⁴¹ В.Э.: Потому, что мы же пользуемся этими мозговыми программами каждый день.

⁴² В.Э.: Вот, к каким нелепым выводам Пенроуза привела математическая парадигма и машины Тьюринга, не давшие ему познать настоящих алгоритмов настоящих программ!

⁴³ 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».

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.⁴⁴ Прежде всего следует решить вопрос об алгоритмической природе внешнего фактора. Может ли реакция внешнего окружения⁴⁵ 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 с тем, 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 всё 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 крите-

⁴⁴ В.Э.: Всё, конечно, зависит от того, какой смысл вкладывать в те или иные слова, но при том смысле, какой мог бы вкладываться в них обычно в программистской среде, это определение Пенроуза неправильно, т.е. оно определяет НЕ самопрограммирование (используемое живыми существами в своей деятельности). Чтобы это определение описывало действительность, его надо переделать следующим образом: «Вообразим, что наша самопрограммирующаяся система помещена в некоторое внешнее окружение, причем в каждый данный момент она должна составить себе программы своего поведения на ближайшее и более отдаленное время. Эти программы непрерывно составляются на основе информации о внешнем окружении и о состоянии самой системы. При этом накапливаются данные о результатах выполнения предыдущих программ и учитываются при составлении следующих. Перед выполнением программ производится симулирование их выполнения для определения и оценки возможных последствий такого выполнения». Вот, это уже более точное описание действительного положения вещей.

⁴⁵ В.Э.: У Пенроуза всё «вывернуто наизнанку»: реагирует как раз система, а не внешнее окружение. (Так ставил бы задачу нормальный программист).

риям, 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 $\mathcal{A}$ или $\mathcal{B}$, 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 окружения⁴⁶ 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 $\mathcal{A}$ или $\mathcal{B}$, 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 5×10^9 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

⁴⁶ В.Э.: При чем тут вообще «численное моделирование окружения»? В разбираемом примере «моделирует» (т.е. строит модели внешнего мира) только самопрограммирующаяся система. Очевидно, что ее модели в любом случае – приближенные (глаз не видит молекул и т.д.). Если модели, по которым ориентируется система в своей деятельности, и без того приближенные, то какое системе дело, может или не может Пенроуз «численно моделировать окружение»?

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, не считая 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 $\mathcal{C}$. Ибо единственной разумной причиной усомниться в возможной справедливости точки зрения $\mathcal{C}$ 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 $\mathcal{C}$. 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 изолированной 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.⁴⁷) 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-то 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

⁴⁷ Популярно об этих исследованиях рассказано в [153] и [337].

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.⁴⁸

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 квантово-механического 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 наши 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.⁴⁹ 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

⁴⁸ В.Э.: Всё перечисленное, однако, не включает четко и ясно поставленной задачи самопрограммирования. Такое впечатление, что – кроме меня – никто в мире не представляет, что это такое. (А самопрограммирование похоже на распускание почки весной – похоже на деление клеток и рост организма. У Природы один почерк – и там, и здесь – и в соматике, и в психике; и в «хардвере» и в «софтвере»).

⁴⁹ Из классической теории фон Неймана и Моргенштерна (1944).

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 всё 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 псевдослучайных 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 всё 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 свою 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 некоторой 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 та самая «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 всё 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 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 им 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. Или 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 всё меньше и меньше ошибок. С течением времени учителя и физическое окружение будут становиться для робота всё менее необходимыми – возможно, в конечном счете, окажутся и вовсе ненужными, – и при формировании своих математических суждений он будет всё 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.⁵⁰ Они

⁵⁰ В.Э.: Нет не так. Всё наоборот! Именно «неопровержимая истина» является «первоначальной установкой» всякого компьютера. Еще до обладания какими бы то ни было зачатками интеллекта любой компьютер действует так, будто вся информация, с которой он работает, – «абсолютно достоверна». Так работают и насекомые – эти наиболее примитивные самопрограммирующиеся автоматы; так работают и животные, например, вороны, кидавшиеся на Конрада Лоренца, когда он держал в руках черную тряпку: для птиц «абсолютно достоверной» была информация, что он – хищник, схвативший одного из воронов. А чтобы начать сомневаться, система должна встроить в себя специальный аппарат, взвешивающий: а так ли это – или не так? (Вообще роботы будут рассуждать совсем не так – абсолютно не так, как это здесь описывает Пенроуз. И математические знания их будут начинаться с изучения программы N натуральных чисел, и подобных программ, – а до разглагольствований о формальных системах они дойдут только в том случае, если совершат те же ошибки, что совершали математики-люди, а скорее всего – не совершат этих

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. 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 $\mathcal{B}$, 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 $\mathcal{B}$, так и $\mathcal{A}$. 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 $\mathcal{B}$ не отличается принципиально от $\mathcal{A}$ в том, что касается ограничений, налагаемых на возможную манеру поведения робота, однако сторонники $\mathcal{B}$, 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, оценивающей обоснованность того или иного математического рассуждения. Подобное человеческое восприятие предполагает всё же некоторое понимание смысла затронутых математических концепций. Согласно точке зрения $\mathcal{A}$, во всем этом нет ничего, выходящего за рамки некоторого свойства вычисления, связанного с понятием «смысла», тогда как $\mathcal{B}$ рассматривает смысл в качестве семантического аспекта мышления и не допускает возможности его описания в чисто вычислительных терминах. В этом мы согласны с точкой зрения $\mathcal{B}$ и отнюдь не ожидаем от нашего робота способности действительно ощущать тонкие семантические различия. Таким образом, сторонники $\mathcal{B}$, возможно, менее (нежели сторонники $\mathcal{A}$) 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 $\mathcal{B}$ будет существенно легче обратить в приверженцев $\mathcal{C}$, чем сторонников $\mathcal{A}$; впрочем, для нашего дальнейшего исследования разница между $\mathcal{A}$ и $\mathcal{B}$ 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 ☆) следует считать неопровержимо истинными – согласно собственным критериям

робота. О допустимости ошибочного присвоения роботом статуса ☆ – пусть им 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 – некоторые 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 всё 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 всё 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 всё, что ему нужно для развития в смысле внешних естественных факторов, – не забывая при этом о том, 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 его 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 *A* или *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 был бы передавать 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 невычис-

лимого. А тем, кто уже и так придерживается точек зрения $\mathcal{C}$ или $\mathcal{D}$ (эти последние, несомненно, склонны, скорее, поверить в возможность существования упомянутой выше невычислимой «сущности», передаваемой роботу в силу одного лишь человеческого происхождения учителя), все эти 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 $\star$ -уровня окажутся, тем не менее, непознаваемыми 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 $\star$ -утверждения) может оказаться 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 $Q(M)$, при этом теоремами такой системы станут следующие положения: (i) $\star$ -утверждения, непосредственно следующие из применения упомянутых механизмов, и (ii) любые положения, выводимые из этих $\star$ -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 $Q(M)$ в силу того простого факта, что процедура $Q(M)$, посредством которой из набора механизмов M получают, одно за другим, необходимые $\star$ -утверждения, является процедурой вычислительной (пусть на практике и весьма громоздкой). Отметим, что определяемая таким образом процедура $Q(M)$ будет генерировать утверждения группы (i), однако вовсе не обязательно все положения группы (ii) (поскольку можно допустить, что нашему роботу, по всей вероятности, попросту надоеет тупо выводить все логические следствия из вырабатываемых им $\star$ -теорем). Таким образом, процедура $Q(M)$ не эквивалентна в точности формальной системе $Q(M)$, однако различие между ними не существенно. К тому же ничто не мешает нам при желании получить из процедуры $Q(M)$ другую процедуру — такую, например, которая будет эквивалентна $Q(M)$.

Далее, для интерпретации формальной системы $Q(M)$ необходимо каким-то образом устроить так, чтобы на всем протяжении развития робота статус $\star$ 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 $\star$ будет иметь совершенно иное значение (либо вовсе окажется бессмысленным). Для того, чтобы определение формальной системы $Q(M)$ на языке робота согласовывалось с нашим ее определением, необходимо в процессе обучения робота (например, учителем-человеком) проследить за тем, чтобы присваиваемое символу $\star$ 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, Π_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 $Q(M)$ также должны быть познаваемыми. Более того, и всякую теорему, выводимую в рамках системы $Q(M)$, 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 $Q(M)$.

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 псевдослучайного (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 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 $Q(M)$. 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 $\mathcal{A}$ и $\mathcal{B}$, 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 $\mathcal{B}$). Иначе говоря, относительно любой заданной (достаточно обширной) формальной системы N робот должен оказаться в силах неопровержимо установить тот факт, что из обоснованности системы N следует истинность его гёделевского⁵¹ утверждения $G(N)$, а также то, что утверждение $G(N)$ не является теоремой системы N . В частности, робот сможет установить, что из обоснованности системы $Q(M)$ неопровержимо следует истинность утверждения $G(Q(M))$; эта же обоснованность предполагает, что утверждение $G(Q(M))$ не является теоремой системы $Q(M)$.

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

⁵¹ В ранних изданиях этой книги вместо обозначения $G(F)$ в оставшейся части главы 3 использовалось обозначение $\Omega(F)$. Однако $G(F)$, на мой взгляд, представляется в данном случае более уместным (см. также §2.8 и с. 160).

убеждений действительно эквивалентна некоей формальной системе $Q(M)$. 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 M , что автоматически означает, что система неопровержимых убеждений робота является полным эквивалентом системы $Q(M)$. Если бы робот вдруг твердо поверил в то, что все его убеждения укладываются в рамки системы $Q(M)$, то тогда ему пришлось бы поверить и в обоснованность этой самой системы $Q(M)$. Соответственно, ему также пришлось бы одновременно поверить и в истинность утверждения $G(Q(M))$ 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 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 $G(Q(M))$), которые роботу оказываются не по силам, хотя изначально предполагалось, что способности робота будут равны способностям человека (или даже превысят их).

* * *

2010.09.04 14:43 суббота

В.Э.: Ну, здесь, пожалуй, нам следует остановиться и оценить происходящее с более общих позиций, а следовать за Пенроузом по всем его мелким тропинкам рассуждений нет уже особого смысла, так как ясно, что у него глобально всё неверно.

Итак, ядром всей установки Пенроуза является теорема Гёделя, а стержень рассуждений: если в основе деятельности робота лежит вычислительная процедура F , то она эквивалентна формальной системе $Q(M)$, а в этой системе появляется теорема Гёделя $G(Q(M))$, которая тогда и всё разрушает и начинаются всякие ужасы...

Самым фундаментальным заблуждением Пенроуза является то, что он полагает, будто вся та среда идей, в которой он находится и рассуждает в этой главе и в которой заставляет действовать также и своего робота, – будто вся эта среда (с формальными системами, теоремами Гёделя и т.д.) является частью математики, принадлежит ей. Но это не так.

Математика – это наука, которая в теоретическом плане зародилась в Древней Греции, а в практическом плане еще раньше на Древнем Востоке. Это была наука о потенциальных продуктах мозговых программ, хотя ее создатели, разумеется, не могли ТАК определять ее предмет и определили ее как науку об объектах Платонова мира идей. Своих вершин эта наука достигла в дифференциальном и интегральном исчислении; образцы описаний математических истин даны в четырех факсимилах в Предисловии к этому тому. За каждой видной там формулой стоит мозговая программа, и разобрать все эти программы не представляет иных трудностей, кроме времени и места.

Если просто разбирать эти программы, изучать их и тем самым устанавливать математические истины, то НИКОГДА невозможно придти ни к «формальным системам», ни к «теореме Гёделя», ни к той среде вообще, в которой здесь рассуждает Пенроуз. Математика во всем этом НЕ НУЖДАЕТСЯ – и поэтому всё это математике НЕ ПРИНАДЛЕЖИТ.

Это построения околоматематические – и порождены они стремлением «обосновать» настоящую математику в условиях, когда действительный ее предмет (мозговые программы) был не известен и не были видны верные пути ее обоснования (в той мере, в какой это «обоснование» вообще было необходимо, – а было оно необходимо в гораздо меньшей степени, чем представлялось тогда, когда на всех давило непонимание «таинственных» истоков математики).

Если мы остаемся в рамках математики (т.е. изучения мозговых программ), то, во-первых, нам вообще не нужна формализация (и аксиоматизация), а во-вторых, теорема Гёделя НЕ ВЕРНА – как мы это увидели в пенроузовском §2.5. (Там, правда, была дана пенроузовская собственная модификация «теоремы Гёделя–Тьюринга», но в принципе аналогичные соображения сделают недействительной и оригинальную «теорему Гёделя»).

Таким образом, если мы остаемся в рамках математики и рассуждаем о мозговых программах так, как вообще нормальные люди могут правдиво рассуждать о программах, то вся околоматематическая пристройка не нужна, а если ее всё-таки построить, то теорема Гёделя в ней несостоятельна.

Несостоятельность теоремы Гёделя (как и аналогичных рассуждений Тьюринга и более ранних оригинальных рассуждений Кантора) вытекает из несостоятельности «диагонального метода» как такового – при рассуждениях о ПРОГРАММАХ он недействителен: программы НЕ ВЕДУТ себя так, как требуется для обоснованного применения этого «метода» и не создают таких бесконечных множеств, о которых можно было бы рассуждать на основе «диагонального метода».

Здесь есть одна тонкость. В принципе можно принять (т.е. постулировать), что «диагональный метод» правильный (т.е. что бесконечные множества не строятся никакими программами, а существуют как «абсолютные бесконечности»). На самом деле современные математики – вслед за Кантором – так и поступают.

Тогда, в той «околоматематической пристройке», в которой принят такой постулат («постулат Кантора»), станут действительными и теоремы Кантора, и Гёделя, и Тьюринга – и всё, о чем говорит Пенроуз.

Но! ТАКАЯ «околоматематическая пристройка» потеряет связь с настоящей, первоначальной математикой. Она отделится от нее и поведет самостоятельное существование.

Первоначальная, настоящая математика создавалась на основе мозговых программ, работающих с предметами реального мира (с их отражениями). Именно поэтому и только поэтому первоначальная, настоящая математика могла (и может) принести такую пользу людям.

А «околоматематическая пристройка», как только она приняла «постулат Кантора» и оторвалась от настоящей математики, эту связь с реальным миром потеряла. И поэтому она НИКОГДА не принесет НИКАКОЙ пользы людям. Это просто бесплодные, пустые спекуляции – не наука, а псевдонаука. (Самолетопоклонники!)

Вот глобальная ориентировка для оценки всех рассуждений Пенроуза в этой главе.

На самом деле Фундаментальное противоречие заключается в следующем. Если рассуждения Пенроуза состоятельны, то они не имеют никакого отношения к математике и к реальному миру (это просто умственные манипуляции «самолетопоклонников»). Если же рассуждения Пенроуза претендуют на связь с математикой и реальным миром, то они несостоятельны.

Продолжаем текст Пенроуза:

* * *

§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 $Q(M)$ 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 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.

(а) Возможно, что робот, принимая в целом утверждение о том, что в основе его конструкции лежит некий набор механизмов M , тем не менее, неизбежно остается неспособен безоговорочно поверить в этот факт.

(б) Возможно, что робот, будучи безоговорочно убежден в истинности каждого отдельного ☆-утверждения в тот момент, когда он его формулирует, всё же сомневается в достоверности полной системы своих ☆-утверждений – соответственно, робот может не верить в то, что формальная система $Q(M)$ и в самом деле лежит в основе всей его системы убеждений в отношении Π_1 -высказываний.

(с) Возможно, что подлинный набор механизмов 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 M в действительности непознаваем.

В последующих девяти разделах представлен ряд веских аргументов, убедительно демонстрирующих, что первые три лазейки ((a), (b) и (с)) 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 $\mathcal{A}$ и $\mathcal{D}$, о которой мы говорили в конце 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 всё 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 невывислимы. 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.⁵² При этом мы на время допускаем, что робот по тем или иным причинам уже отбросил варианты (b), (с) 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 $\mathcal{M}$ гипотезу «В основе математического понимания робота лежит набор механизмов M » и рассмотрим утверждение вида «Такое-то Π_1 -высказывание является следствием из $\mathcal{M}$ ».

Такое утверждение (в том случае, когда робот твердо верит в его истинность) я буду называть $\star_{\mathcal{M}}$ -утверждением. Иначе говоря, под $\star_{\mathcal{M}}$ -утверждениями не обязательно понимаются те Π_1 -высказывания, в истинность которых как таковых неопровержимо верит робот, но те Π_1 -высказывания, которые робот полагает неопровержимо выводимыми из гипотезы $\mathcal{M}$. 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 Π_1 -высказывания, которые робот должен полагать неопровержимыми следствиями из гипотезы $\mathcal{M}$ и которые при этом не являются самыми обыкновенными $\star$ -утверждениями, вовсе не требующими привлечения этой гипотезы? Разумеется, существуют. Как

⁵² В.Э.: Сам Пенроуз и представляет собой такого робота! ☺

было отмечено в конце §3.14, истинность Π_1 -высказывания $G(Q(M))$ следует из обоснованности формальной системы $Q(M)$, отсюда же следует и тот факт, что утверждение $G(Q(M))$ не является теоремой системы $Q(M)$. 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 $Q(M)$, будь он действительно сконструирован в соответствии с набором механизмов M , – т.е. что возможность (b)⁵³ он из рассмотрения исключает, – то получается, что наш робот и в самом деле должен твердо верить в то, что обоснованность системы $Q(M)$ является следствием гипотезы $\mathcal{M}$. Таким образом, робот оказывается безоговорочно убежден как в том, что Π_1 -высказывание $G(Q(M))$ следует из гипотезы $\mathcal{M}$, так и в том, что (согласно $\mathcal{M}$) он не способен непосредственно постичь его неопровержимую истинность без привлечения $\mathcal{M}$ (поскольку формальной системе $Q(M)$ оно не принадлежит). Соответственно, утверждение $G(Q(M))$ является $\star_{\mathcal{M}}$ -утверждением, но не $\star$ -утверждением.

Предположим, что формальная система $Q_{\mathcal{M}}(M)$ построена в точности так же, как и система $Q(M)$, с той лишь разницей, что роль, которую при построении системы $Q(M)$ исполняли $\star$ -утверждения, сейчас берут на себя $\star_{\mathcal{M}}$ -утверждения. Иначе говоря, теоремами системы $Q_{\mathcal{M}}(M)$ являются либо (i) сами $\star_{\mathcal{M}}$ -утверждения, либо (ii) положения, выводимые из этих $\star_{\mathcal{M}}$ -утверждений с применением правил элементарной логики (см. §3.13). Точно так же, как робот на основании гипотезы $\mathcal{M}$ согласен с тем, что формальная система $Q(M)$ охватывает все его неопровержимые убеждения относительно истинности Π_1 -высказываний, он будет согласен и с тем, что формальная система $Q_{\mathcal{M}}(M)$ охватывает все его неопровержимые убеждения относительно истинности Π_1 -высказываний, обусловленных гипотезой $\mathcal{M}$.

Далее предложим роботу рассмотреть гёделевское Π_1 -высказывание $G(Q_{\mathcal{M}}(M))$. Робот, несомненно, проникнется неопровержимым убеждением в том, что это Π_1 -высказывание является следствием из обоснованности системы $Q_{\mathcal{M}}(M)$. Он также вполне безоговорочно поверит в то, что обоснованность системы $Q_{\mathcal{M}}(M)$ является следствием гипотезы $\mathcal{M}$, поскольку он согласен с тем, что система $Q_{\mathcal{M}}(M)$ действительно содержит в себе всё, в чем робот неопровержимо убежден в отношении своей способности выводить Π_1 -высказывания, основываясь на гипотезе $\mathcal{M}$. (Он будет рассуждать следующим образом: «Если я принимаю гипотезу $\mathcal{M}$, то я тем самым принимаю и все Π_1 -высказывания, которые порождают систему $Q_{\mathcal{M}}(M)$. Таким образом, я должен согласиться с тем, что система $Q_{\mathcal{M}}(M)$ является обоснованной на основании гипотезы $\mathcal{M}$. Следовательно, на основании всё той же гипотезы, я должен признать и то, что утверждение $G(Q_{\mathcal{M}}(M))$ истинно».)

Однако, поверив (безоговорочно) в то, что гёделевское Π_1 -высказывание $G(Q_{\mathcal{M}}(M))$ является следствием гипотезы $\mathcal{M}$, робот вынужден будет поверить и в то, что утверждение $G(Q_{\mathcal{M}}(M))$ является теоремой формальной системы $Q_{\mathcal{M}}(M)$. А в это он сможет поверить только в том случае, если он полагает систему $Q_{\mathcal{M}}(M)$ необоснованной, – что решительно противоречит принятию им гипотезы $\mathcal{M}$.

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 $\mathcal{M}$, равно как и в определении $\star_{\mathcal{M}}$ -утверждения, некоторую неоднозначность. Смею вас уверить, что подобное утверждение, будучи Π_1 -высказыванием, представляет собой в высшей степени определенное математическое утверждение. Можно предположить, что большинство $\star_{\mathcal{M}}$ -утверждений робота окажутся в действительности самыми обыкновенными $\star$ -утверждениями, поскольку маловероятно, что робот при каких угодно обстоятельствах сочтет целесообразным прибегать в своих рассуждениях к самой гипотезе $\mathcal{M}$. Исключением может стать утверждение $G(Q(M))$ о котором говорилось выше, так как в данном случае формальная система $Q(M)$ выступает, с точки зрения робота, в роли гёделевской гипотетической «машины для доказательства теорем» (см. §§3.1 и 3.3). Вооружившись гипотезой $\mathcal{M}$, робот получает доступ к своей собственной «машине для

⁵³ 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.

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 M , а не просто отдельная формальная система $Q(M)$, он может повторить свое рассуждение и перейти от системы $Q(M)$ к системе $Q_{\mathcal{M}}(M)$, обоснованность которой он по-прежнему полагает простым следствием из гипотезы $\mathcal{M}$. Именно это и приводит его в конечном итоге к противоречию (чего мы, собственно, и добивались). (См. также §3.24, где мы продолжим рассмотрение системы $Q_{\mathcal{M}}(M)$ 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 постижимого 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 $Q(M)$ на основе механизмов M , и в самом деле охватывает всю совокупность Π_1 -высказываний, в истинность которых он неопровержимо верит, а также, соответственно, с тем, что формальная система $Q_{\mathcal{M}}(M)$ охватывает всю совокупность Π_1 -высказываний, которые, как он неопровержимо верит, следуют из гипотезы $\mathcal{M}$. Кроме того, если мы хотим, чтобы робот смог построить собственную потенциально непротиворечивую систему математических убеждений, следует ввести в набор механизмов M какие-либо подлинно случайные составляющие.

Эти последние оговорки мы рассмотрим в последующих разделах (§§3.17–3.22). Вопрос о введении в набор механизмов M возможных случайных элементов (вариант (с)) 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 M – формальная система $Q(M)$ корректным образом включает в себя всю систему его математических убеждений в отношении Π_1 -высказываний (равно как и с соответствующим предположением для системы $Q_{\mathcal{M}}(M)$)? Такое согласие подразумевает, прежде всего, что робот верит в обоснованность системы $Q(M)$ – т.е. в то, что все Π_1 -высказывания, являющиеся $\star$ -утверждениями, действительно истинны. Наши рассуждения требуют также, чтобы всякое Π_1 -высказывание, в истинность которого робот в состоянии безоговорочно поверить, являлось непременно теоремой системы $Q(M)$ (т.е. чтобы в рамках системы $Q(M)$ 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 $Q(M)$ действительно играла такую универсальную роль в отношении потенциальных способностей робота, связанных с Π_1 -высказываниями, а лишь то, чтобы она была достаточно обширна для того, чтобы допускать применение гёделевского доказательства к самой себе (и, соответственно, к системе $Q_{\mathcal{M}}(M)$). Позднее мы увидим, что необходимость в таком применении возникает лишь в случае некоторых конечных систем Π_1 -высказываний.

Таким образом, мы – как, собственно, и робот – должны учитывать возможность того, что некоторые из $\star$ -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 Π_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 Π_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 $\star$ -утверждения, вообще говоря, полагаться нельзя, пусть даже он и удостоил их самолично статуса $\star$. Впоследствии робот может исправить свою ошибку, однако ошибка-то уже сделана. Каким образом это обстоятельство отразится на нашем выводе относительно обоснованности формальной системы $Q(M)$? Очевидно, что система $Q(M)$ не является целиком и полностью обоснованной, не «воспринимает» ее как таковую и робот, так что его гёделевскому предположению $G(Q(M))$ 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 $\mathcal{B}$, нежели $\mathcal{A}$), 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 $\star$ -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 $\star$ -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 $Q(M)$, нам, судя по всему, придется полагаться лишь на доступные $\star$ -утверждения, в достоверности которых мы не можем быть полностью уверены.

Не здесь ли проходит возможная операционная граница между точками зрения $\mathcal{A}$ и $\mathcal{B}$? Не исключено, что так оно и есть; хотя позиции $\mathcal{A}$ и $\mathcal{B}$ 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 $\star$ -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, во всяком случае, чрезвычайно редки. Разумно будет предположить, что всякий раз, когда робот допускает ошибку в одном из своих $\star$ -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, – однако вполне может совершить другие. Упомянутые факторы могут приноситься 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 нечслучайное (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 и очень большим, и всё 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 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 соображений 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 М – 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 всё 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 обозначаю буквой М), 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) P_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 P_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 P_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 P_1 -высказывания.

Понятию «сложности» применительно к P_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. 191). Итак, степенью сложности P_1 -высказывания, утверждающего незавершаемость вычисления $T_m(n)$ машины Тьюринга, мы будем полагать число p знаков в двоичном представлении большого 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» – просто на всякий случай – сформулировать его опровержение. В отсутствие числа L

такое $\star$ -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 $\star$ -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 $T(p)$, $L(p)$ и $N(p)$ вряд ли окажутся чем-то из ряда вон выходящим. Предположив, что всё 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 $Q'(M)$ (также познаваемую), теоремами которой являются те самые безошибочные $\star$ -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 $\mathcal{B}$ 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 $\star$ -утверждения действительно безошибочны, исходя из допущения, что именно набором механизмов M и определяется его поведение (гипотеза $\mathcal{M}$ из §3.16). До сих пор, в данном разделе, мы занимались исключительно устранением ошибок в $\star$ -утверждениях робота. Однако, на самом деле, ввиду представленного в §3.16 фундаментального противоречия, нас интересует устранение ошибок в его $\star_{\mathcal{M}}$ -утверждениях, т.е. в тех Π_1 -высказываниях, что по неопровержимой убежденности робота следуют из гипотезы $\mathcal{M}$. Поскольку принятие роботами формальной системы $Q'(M)$ в любом случае обусловлено гипотезой $\mathcal{M}$, мы вполне можем предложить им для обдумывания и более обширную формальную систему $Q'_{\mathcal{M}}(M)$, определяемую аналогично формальной системе $Q_{\mathcal{M}}(M)$ из §3.16. Под $Q'_{\mathcal{M}}(M)$ в данном случае понимается формальная система, построенная из $\star_{\mathcal{M}}$ -утверждений, «безошибочность» которых установлена в соответствии с вышеописанными критериями T , L и N . В частности, утверждение «утверждение $G(Q'_{\mathcal{M}}(M))$ истинно» считается здесь безошибочным $\star_{\mathcal{M}}$ -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 (вкуче 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 $\star_{\mathcal{M}}$ - или $\star$ -утверждения? В конце концов, приведенные выше рассуждения будут иметь смысл лишь в том случае, если нам удастся исключить абсолютно все ошибочные $\star_{\mathcal{M}}$ -утверждения (или $\star$ -утверждения) в отношении Π_1 -высказываний. Окончательно и бесповоротно удостовериться в истинности утверждения $G(Q'_{\mathcal{M}}(M))$ нам (и роботам) поможет обоснованность формальной системы $Q'_{\mathcal{M}}(M)$ (обусловленная гипотезой $\mathcal{M}$). Эта самая обоснованность подразумевает, что система $Q'_{\mathcal{M}}(M)$ ни в коем случае не может содержать таких $\star_{\mathcal{M}}$ -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 $\star_{\mathcal{M}}$ -утверждений

Есть, впрочем, возможность именно эту конкретную проблему разрешить и сузить область рассмотрения до конечного множества различных $\star_{\mathcal{M}}$ -утверждений. Само доказательство

несколько громоздко, однако основная идея заключается в том, что нам необходимо рассматривать только те Π_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 Π_1 -высказывания. «Максимальная длина» задается неким числом c , которое можно определить из степени сложности правил, определяющих формальную систему $Q'_m(M)$. 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» Π_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 порекомендую просто-напросто пропустить весь этот материал.

Нам понадобится несколько модифицировать формальную систему $Q'_m(M)$, приведя ее к виду $Q'_m(M, c)$ – для краткости я буду обозначать ее просто как $Q(c)$ (отброшенные обозначения в данной ситуации несущественны и лишь добавляют путаницы и громоздкости). Формальная система $Q(c)$ определяется следующим образом: при построении этой системы допускается принимать в качестве «безошибочных» только те $\star_m$ -утверждения, степень сложности которых (задаваемая описанным выше числом ρ) меньше c , где c есть некоторое должным образом выбранное число, подробнее о котором я расскажу чуть ниже. Для «безошибочных» $\star_m$ -утверждений, удовлетворяющих неравенству $\rho < c$, я буду использовать обозначение « $\sqrt{\text{краткие}} \star_m$ -утверждения». Как и прежде, множество действительных теорем формальной системы $Q(c)$ будет включать в себя не только $\sqrt{\text{краткие}} \star_m$ -утверждения, но также и утверждения, получаемые из $\sqrt{\text{кратких}} \star_m$ -утверждений посредством стандартных логических операций (позаимствованных, скажем, из исчисления предикатов). Хотя количество теорем системы $Q(c)$ бесконечно, все они выводятся с помощью обыкновенных логических операций из конечного множества $\sqrt{\text{кратких}} \star_m$ -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 ρ). Таким образом, формальная система $Q(c)$ задается лишь четырьмя постоянными c , T , L , N и общей системой механизмов 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 $G(H)$ для формальной системы H является Π_1 -высказывание, степень сложности которого должна лишь на сравнительно малую величину превышать степень сложности самой системы H , причем эту величину можно определить точно.

Конкретности ради я позволю себе некоторое нарушение системы обозначений и буду вкладывать в запись $G(H)$ некий особый смысл, который может и не совпасть в точности с определением, данным в §2.8. В формальной системе H нас интересует лишь ее способность доказывать Π_1 -высказывания. В силу этой своей способности система H дает нам алгебраическую процедуру A , с помощью которой мы можем в точности установить (на основании завершения выполнения A) справедливость тех Π_1 -высказываний, формулировка которых допускается правилами системы H . А под Π_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 H возможно установить справедливость того самого Π_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 H и истинное Π_1 -высказывание, которое системе H оказывается «не по зубам». Именно это Π_1 -высказывание я буду теперь обозначать через $G(H)$. Оно существенно эквивалентно (при условии достаточной обширности) действительному

утверждению «система H непротиворечива», хотя в некоторых деталях эти два утверждения могут и не совпадать (см. §2.8).

Пусть α 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 α , где $A = T_\alpha$. Тогда, согласно построению, представленному в явном виде в Приложении A, находим, что степень сложности η утверждения $G(H)$ удовлетворяет неравенству $\eta < \alpha + 210 \log_2 (\alpha + 336)$. Для нужд настоящего рассуждения мы можем определить степень сложности формальной системы H как равную степени сложности процедуры A т.е. числу α . Приняв такое определение, мы видим, что «излишек» сложности, связанный с переходом от H к $G(H)$ оказывается еще меньше, чем и без того относительно крохотная величина $210 \log_2 (\alpha + 336)$.

Далее нам предстоит показать, что если $H = Q(c)$ при достаточно большом c , то $\eta < c$. Отсюда, соответственно, следует, что и Π_1 -высказывание $G(Q(c))$ должно оказаться в пределах досягаемости системы $Q(c)$ при условии, что роботы принимают $G(Q(c))$ с $\star$ -убежденностью. Доказав, что $c > \gamma + 210 \log_2 (\gamma + 336)$, мы докажем и то, что $\gamma < c$; буквой γ мы обозначили значение α при $H = Q(c)$. Единственная возможная сложность здесь обусловлена тем обстоятельством, что сама величина γ зависит от c , хотя и не обязательно очень сильно. Эта зависимость γ от c имеет две различных причины. Во-первых, число c являет собой явный предел степени сложности тех Π_1 -высказываний, которые в определении формальной системы $Q(c)$ называются «безошибочными $\star_{\mathcal{M}}$ -утверждениями», вторая же причина происходит из того факта, что система $Q(c)$ явным образом обусловлена выбором чисел T , L и N и можно предположить, что для принятия в качестве «безошибочного» $\star_{\mathcal{M}}$ -утверждения большей сложности необходимы какие-то более жесткие критерии.

Относительно первой причины зависимости γ 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 γ 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 $2^{2^{s^2}}$ 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 $Q(c)$ записывать Q^* . Итак, Q^* есть формальная система, теоремами которой являются все математические высказывания, какие можно вывести из конечного количества $\sqrt{k}$ кратких $\star_{\mathcal{M}}$ -утверждений, используя стандартные логические правила (исчисление предикатов). Количество этих $\star_{\mathcal{M}}$ -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 $\star_{\mathcal{M}}$ -убежденностью, то они, несомненно, $\star_{\mathcal{M}}$ -заклучают, что гёделевское предположение $G(Q^*)$ также истинно на основании гипотезы $\mathcal{M}$, поскольку является Π_1 -высказыванием меньшей, нежели c , сложности. Рассуждение для получения утверждения $G(Q^*)$ из $\star_{\mathcal{M}}$ -убежденности в обоснованности формальной системы Q^* достаточно просто (в сущности, я его уже привел), так что с присвоением этому утверждению статуса $\star_{\mathcal{M}}$ проблем возникнуть не должно. То есть само $G(Q^*)$ также должно быть теоремой системы Q^* . Это, однако, противоречит убежденности роботов в обоснованности Q^* . Таким образом, упомянутая убежденность (при условии справедливости гипотезы $\mathcal{M}$ 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, 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 приходим 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,⁵⁴ т.е. увеличение числа T выше определенного значения на деле увеличивает и вероятность ошибки в $\star_{III}$ -утверждениях. С другой стороны, если слишком большим сделать N (или L), то возникает риск исключить вообще все $\star_{III}$ -утверждения из-за существующего в сообществе меньшинства «глупых» роботов, раздражающихся время от времени произвольными $\star_{III}$ -утверждениями, которые в данном случае не перекроются необходимым количеством $\star$ -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 $\star$ присваивался утверждениям только с одобрения всего сообщества роботов в целом.

Существует и много других возможностей улучшения качества $\star_{III}$ -утверждений или исключения ошибочных утверждений из общего (конечного) их числа. Кого-то, возможно, беспокоит тот факт, что, несмотря на установление предела c сложности P_1 -высказываний, ограничивающего общее количество кандидатов на $\star$ - или $\star_{III}$ -статус до некоторой конечной величины, эта величина окажется всё же чрезвычайно огромной (будучи экспоненциально зависимой от c), вследствие чего становится весьма сложно однозначно удостовериться, что исключены все возможные ошибочные $\star_{III}$ -утверждения. В самом деле, никакого ограничения не задается в рамках нашей модели на количество «робото-вычислений», необходимых для получения удовлетворительного $\star_{III}$ -доказательства какого-либо из P_1 -высказываний. Следует ввести четкое правило: чем длиннее в таком доказательстве цепь рассуждений, тем более жесткие критерии применяются при решении вопроса о присвоении ему $\star_{III}$ -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 $\star_{III}$ -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 четкого сформулированного и вычислимого условия, достаточного для устранения всех ошибочных $\star_{III}$ -

⁵⁴ В.Э.: Да-а – это мы наблюдаем у математиков.

утверждений. В результате мы приходим к строгому выводу: никакие познаваемые механизмы, пусть и снабженные какими угодно вычислительными «подпорками», не способны воспроизвести корректное математическое умозаключение человека.

Мы рассматривали ☆-утверждения, которые, оказавшись по той или иной причине ошибочными, в принципе исправимы самими роботами, – пусть даже в каком-то конкретном экземпляре модели робота 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 М (неизменных нисходящих, «самоусовершенствующихся» восходящих либо и тех, и других в какой угодно пропорции) не может обуславливать поведение нашего сообщества роботов, равно как и отдельных его членов, – если исходить из допущения, что роботы способны достичь человеческого уровня математического понимания.⁵⁵ 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,⁵⁶ 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 невывислительное 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 всё же, безотносительно к упомянутой возможности, хаос может предоставить нам лишь очень сомнительный способ обойти неутешительное заключение, к которому мы пришли в предыдущем разделе. 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 М. 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 всё 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 возмож-

⁵⁵ В.Э.: То есть – поверить в правильность доказательства Гёделя (хе-хе-хе...)

⁵⁶ См. [153], [337].

ных альтернативных робото-историй. Однако само хаотическое поведение нам просто-напросто придется вычислять – в чем нет ничего странного: на практике, в математических примерах, хаотическое поведение обыкновенно и вычисляется на компьютере. Ансамбль возможных альтернатив окажется в данном случае не таким большим, каким он мог бы быть, допустим 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 $\star_{III}$ -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 $\star_{III}$ -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 такого 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(Q) – в роли незавершающегося вычисления $C_k(k)$.⁵⁷ То есть к чтению нижеследующего материала можно переходить сразу после §2.5 без какого бы то ни было ущерба для понимания.]

Альберт Император⁵⁸ 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.

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?

⁵⁷ В.Э.: Которое, как мы убедились, не существует.

⁵⁸ В.Э.: Вообще это надо было переводить как «Альберт Король», раз уж название первой книги Пенроуза «The Emperor's New Mind» перевели как «Новый Разум Короля» (НРК); ведь ясно же, что это тот же самый *Emperor*, что и там – и ассоциация там была со сказкой Андерсена, которая по-английски называется «The Emperor's New Clothes», а по-русски «Новый Наряд Короля» (король, как известно, на самом деле был голый). Имя *Emperor*-а, Альберт, видимо, взято от *Albert Magnus* (Альберт Великий – средневековый философ). Пенроуз, таким образом, во всю издевается над теми, кто считает ИИ реализуемым на компьютерах. (Вот, за это он и получил от меня ту картинку с подписью, которая стоит на титульном листе настоящего тома). Или можно было не переводить вовсе и оставить как английскую фамилию: Альберт Эмперор.

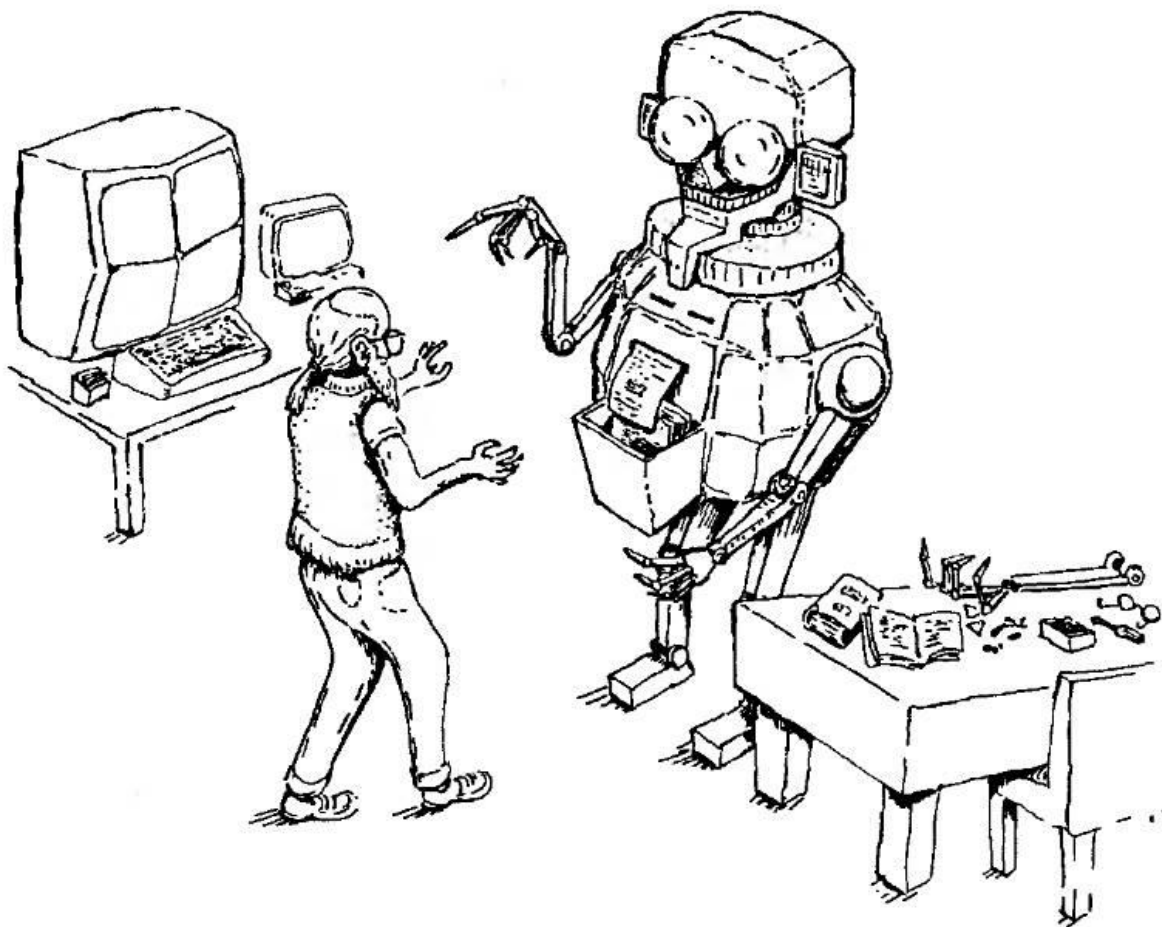


Рис. 3.2. Альберт Император и Математический Интеллектуальный Киберкомплекс.

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 здоровой 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 всё 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, опасаясь, 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. Всё 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 большей 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 всё 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 уважаемых 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 генерации 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. Всё развитие вашего сообщества роботов представляет собой выполнение некоего невероятно 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 всё 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 $\star$ -утверждениями, которые являются P_1 -высказываниями. Ты помнишь, что такое P_1 -высказывание?

М.И.К.: Мне, разумеется, прекрасно известно определение P_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 $\star$ -утверждаемые P_1 -высказывания, через $Q(M)$ или, для краткости, просто буквой Q . Логичным будет предположить, что должно существовать некое математическое утверждение гёделевского типа – также P_1 -высказывание, обозначим⁵⁹ его через $G(Q)$ – причем истинность $G(Q)$ является следствием утверждения, что вы, роботы, никогда не допускаете ошибок в отношении P_1 -высказываний, которым вы присваиваете статус $\star$.

М.И.К.: Да; тут ты, надо полагать, тоже прав... гм.

А.И.: И утверждение $G(Q)$ должно быть истинным, поскольку вы, роботы, никогда не ошибаетесь в ваших $\star$ -утверждениях.

М.И.К.: Разумеется.

А.И.: Минуточку... отсюда также следует, что роботы должны быть неспособны установить истинность утверждения $G(Q)$ по крайней мере, с $\star$ -уверенностью.

М.И.К.: Тот факт, что мы, роботы, были изначально сконструированы в соответствии с набором механизмов M вкупе с тем фактом, что наши $\star$ -утверждения, касающиеся P_1 -высказываний, никогда не бывают ошибочными, и в самом деле имеет очевидное и неопровержимое следствие, заключающееся в том, что P_1 -высказывание $\Omega(Q)$ должно быть истинным. Полагаю, ты думаешь, что я наверняка смогу убедить СМISР присвоить утверждению $G(Q)$ статус $\star$, 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 $\star$ -статуса как раз и заключается в том, что он является гарантией правильности.

Хотя ... невозможно, чтобы они смогли согласиться с утверждением $G(Q)$, так как по самой природе твоего гёделевского построения это утверждение не входит в число тех предположений, истинность которых мы можем установить с $\star$ -уверенностью – при условии, что мы в своих $\star$ -утверждениях действительно не ошибаемся. Полагаю, ты намекаешь на то, что эта несообразность должна посеять в нас какие-то сомнения относительно адекватности наших $\star$ -суждений.

Я, однако, и мысли не допускаю о том, что наши $\star$ -утверждения могут оказаться ложными, особенно если учесть всю тщательность их рассмотрения и предпринимаемые СМISР 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, СМISР 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

⁵⁹ 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.

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 Π_1 -высказывание $G(Q)$ и однозначно установить, что оно истинно – при условии, конечно же, что роботы и в самом деле никогда не ошибаются в своих $\star$ -утверждениях. Вы же не можете быть уверенными в том, что высказывание $G(Q)$ истинно; во всяком случае, вы не можете утверждать этого с той убежденностью, какой, несомненно, потребует СМИСР для присвоения $\star$ -статуса. Это, похоже, дает людям некое фундаментальное преимущество перед роботами, пусть даже только в принципе, а не на практике – существуют такие Π_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 Π_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 Π_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-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 Π_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 Π_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 $\star$ -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 $\star$ -статус. Полагаю, возможно и такое, что члены СМИСР не убеждены безошибочно в том, что присвоение ими $\star$ -статуса неизменно происходит безошибочно. Таким образом, утверждение $G(Q)$ может и не приобрести $\star$ -статуса, и противоречие исчезнет само собой. Заметь себе, это вовсе не означает, что я признаюсь в том, что мы, роботы, намеренно делаем ошибочные $\star$ -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 $\star$ -утвержденного Π_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 Π_1 -высказываний бесконечно много? Мне почему-то вспомнилось об условии ω -непротиворечивости, которое, если не ошибаюсь, имеет какое-то отношение к гёделевскому утверждению $G(Q)$.

М.И.К.: (после едва заметно более продолжительной паузы) Нет, определенно нет. Это никак не связано с тем, что число возможных Π_1 -высказываний бесконечно. Мы можем ограничить рассмотрение только теми Π_1 -высказываниями, которые являются в некотором вполне определенном смысле «краткими», – т.е. такими, что описание машины Тьюринга для каждого из них содержит не более s двоичных знаков, где s 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 $G(Q)$ – неизменна и довольно проста, нет необходимости рассматривать Π_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 Π_1 -высказывания составляют конечное семейство, пусть и весьма многочисленное. Ограничив рассмотрение лишь «краткими» Π_1 -высказываниями, мы получаем некоторую вычислительную процедуру Q^* – той же, в сущности, сложности, что и процедура Q , – которая будет генерировать только такие $\star$ -утверждаемые краткие Π_1 -высказывания. К этой новой процедуре применимы все наши прежние рассуждения. Исходя из заданной процедуры Q^* , мы можем отыскать другое краткое Π_1 -высказывание, которое, разумеется, должно быть истинным – при условии, что истинными являются все $\star$ -утверждаемые краткие Π_1 -высказывания, – однако истинность его невозможно установить с $\star$ -уверенностью. Впрочем, всё 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 Π_1 -высказываний, истинность каждого из которых в отдельности гарантирована, однако никто из вас, ни СМISР, 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 $G(Q^*)$, которая есть следствие истинности всех Π_1 -высказываний из этого самого ряда. Как-то нелогично, не находишь?

М.И.К.: Роботы не могут быть нелогичными. Π_1 -высказывание $G(Q^*)$ является следствием из остальных Π_1 -высказываний только в том случае, если мы действительно были построены в соответствии с механизмами M . Мы не можем гарантировать истинности $G(Q^*)$ 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 Π_1 -высказывания $G(Q^*)$, однако в нашем случае имеется иная неопределенность: мы не можем разделить эту вашу твердолобую уверенность в том, что все ваши $\star$ -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? Вيني нас, если угодно, однако я бы на твоём месте предположил, что некоторый элемент неопределенности в утверждении «все $\star$ -утверждаемые краткие Π_1 -высказывания непременно истинны» всё же присутствует, потому хотя бы, что ты не веришь, что мы при твоём 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 $\star$ -утверждениями (число которых, если помнишь, является конечным), они не составят сколько-нибудь существенной неопределенности в утверждении $G(Q^*)$.

Кроме того, есть еще кое-что, о чем ты, возможно, и не подозреваешь. Нам необходимо рассматривать лишь те $\star$ -утверждения, что удостоверяют истинность того или иного Π_1 -высказывания (более того, краткого Π_1 -высказывания). Не может быть никакого сомнения в том, что разработанные СМISРом тщательнейшие процедуры исключают 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 Π_1 -высказываний, в соответствии с которой СМISР может полагать неопровержимо истинным какое-либо краткое Π_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 твоё утверждение о том, что в основе нашей конструкции лежат именно механизмы М, – а я всё 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 СМISР утверждает $\star$ -статус математического доказательства, были столь вопиюще ложными.

А.И.: Значит, существенной (иначе говоря, избавляющей тебя от необходимости присваивать $\star$ -статус утверждению $G(Q^*)$), чего, как тебе известно, ты сделать не можешь, не признав прежде, что какие-то из прочих $\star$ -утвержденных кратких Π_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 М. А раз ты не можешь поверить в то, о чем мы знаем, ты не можешь и доказать истинность утверждения $G(Q^*)$, тогда как мы можем это сделать, опираясь на непогрешимость твоих же $\star$ -утверждений, в каковой ты так настойчиво меня убеждаешь.

Я тут припомнил еще кое-что из той занятой древней книжки, о которой я тебе говорил. 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 М, или нет, достаточно, чтобы ты просто допустил, что такое логически возможно. Как же там было... да, вспомнил. Основная идея сводится к следующему: СМISРу необходимо будет учредить еще одну категорию для утверждений, в истинности которых они не так безоговорочно убеждены, – скажем, $\star_{///}$ -утверждений, – но которые они будут рассматривать как неопровержимые следствия из допущения, что все роботы построены в соответствии с набором механизмов М. Эти $\star_{///}$ -утверждения будут, разумеется, включать в себя и все первоначальные $\star$ -утверждения, а также все те утверждения, которые роботы смогут вывести, исходя из допущения, что их действиями управляют именно механизмы М. 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 $\star_{///}$ -утверждений непременно войдет утверждение $G(Q^*)$, а также любое Π_1 -высказывание, которое можно вывести из $G(Q^*)$ и из $\star$ -утверждений с помощью правил элементарной логики. Однако, кроме этих, там будут и другие утверждения. Идея такова, что знание правил М дает возможность получить новую алгоритмическую процедуру $Q^{*///}$, которая будет генерировать только такие (разумеется, краткие) $\star_{///}$ -утверждения (а также логические следствия из них), истинность которые СМISР сможет подтвердить, исходя из допущения, что в основе конструкции роботов лежат именно правила М.

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 $Q^{*///}$... Да, а еще я предвосхитил твой следующий шаг: я составил также гёделевское предположение для этого алгоритма, Π_1 -высказывание $G(Q^{*///})$. 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 $G(Q^{*///})$ – неопровержимо ли оно истинно?

М.И.К.: Неопровержимо истинно? Что ты имеешь в виду? А, понятно... СМISР подтвердит истинность – неопровержимую истинность, если угодно, – высказывания $G(Q^{*///})$, но только при допущении, что в основе конструкции роботов лежат правила М, – а это допущение,

как тебе известно, я нахожу всё более и более сомнительным. Дело в том, что истинность «высказывания $G(Q^*_m)$ » в точности следует из следующего утверждения: «Все краткие Π_1 -высказывания, которые СМISР готов признать неопровержимо истинными, исходя из допущения, что роботы построены в соответствии с правилами М, являются истинными». Так что я не знаю, истинно ли на самом деле высказывание $G(Q^*_m)$. Это зависит от того, справедливо твое сомнительное утверждение или нет.

А.И.: Ясно. Значит, твои слова надо понимать так, что ты (вместе со СМISРом) готов признать – без каких бы то ни было оговорок, – что истинность высказывания $G(Q^*_m)$ следует из допущения, что роботы построены в соответствии с правилами М.

М.И.К.: Разумеется.

А.И.: Тогда получается, что Π_1 -высказывание $G(Q^*_m)$ должно быть $\star_m$ -утверждением.

М.И.К.: Ну коне... гм... что? Ах да, разумеется, ты прав. Однако по самому своему определению, $G(Q^*_m)$ не может само быть $\star_m$ -утверждением, разве что, по меньшей мере, одно из $\star_m$ -утверждений является в действительности ложным. Да ... это только подтверждает то, о чем я тебе всё это время говорю; теперь я могу, наконец, совершенно определенно заявить, что правила или механизмы М 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 всё 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 М 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 всё еще совсем не убежден, что ты был до конца честен со мной в том, что ты говорил мне о механизмах М. 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 *chaos/urpn-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 всё равно не понимаю, что это в практическом смысле меняет? Почти наверняка вычислительная процедура Q (или Q^* , или Q^*_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 М, включая в себя все процедуры обучения и все вычисления в рамках «естественного отбора», благодаря которым мы получили в конце концов наинумнейших роботов в мире.

Однако, полагаю, нельзя полностью исключить возможности (крайне маловероятной, 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, всё сразу стало ясно.

Я и раньше замечал, насколько мои способности к рассуждению превосходят способности даже умнейших из людей. И уж совершенно невозможно для меня смириться с возможностью существования некоего Π_1 -высказывания – такого, например, как $G(Q^*)$ – 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 тех якобы «случайных» элементов, внедрил в нас могучую Алгоритмическую Процедуру, которая кардинально превосходит всё то,

чего способны достичь алгоритм человеческий или гёделевская «машина для доказательства теорем». Мы, роботы, олицетворяем собой фундаментальный прорыв, и человеку со всеми его достижениями 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 пошло 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 Π_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.⁶⁰ 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,

⁶⁰ В оригинале речь идет лишь об английском языке, однако, как нам представляется, английский язык в этом отношении отнюдь не одинок. *Прим. перев.*

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 отношении. Например, определение $\star_{III}$ -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 всё же в этом смысле самоотносимыми $\star_{III}$ -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 P_1 -высказывания P_0 . Робот, возможно, окажется неспособен непосредственно установить, является ли высказывание P_0 в действительности истинным или нет, однако он может обратить внимание на то, что истинность P_0 следует из предположения, что истинным является каждый член некоторого вполне определенного бесконечного класса P_1 -высказываний S_0 (пусть это будут, скажем, теоремы формальной системы $Q(M)$ или $Q_{III}(M)$, 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 вычисления осуществляется построение некоторой модели сообщества математических роботов, а результат S_0 представляет собой семейство P_1 -высказываний, $\star$ -утверждаемых этими самыми моделируемыми роботами. Если механизмы, лежащие в основе этого сообщества роботов, совпадают с набором механизмов M , то высказывание P_0 представляет собой пример $\star_{III}$ -утверждения. А наш робот придет к выводу, что если он сам построен в соответствии с набором механизмов M , то высказывание P_0 также должно быть истинным.

Рассмотрим случай с более тонким $\star_{III}$ -утверждением (обозначим его P_1): робот отмечает, что истинность P_1 является следствием истинности всех членов другого класса P_1 -высказываний (например, S_1), который можно получить из результата того же самого вычисления, моделирующего сообщество роботов (на основе механизмов M), только на этот раз существенная часть результата состоит из, скажем, тех P_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 неперенное следствие допущения, что он построен в соответствии с механизмами M ? Его рассуждение будет выглядеть приблизительно так: «Если в основе моей конструкции лежат механизмы 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 $\star_{III}$ -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 M . Эту цепочку можно, очевидно, продолжать и дальше, приводя $\star_{III}$ -утверждения всё большей и большей тонкости (P_n), истинность которых будет следовать из допущения истинности всех членов классов $S_0, S_1, S_2, S_3, \dots$ и так далее, включая и классы с индексами более высокого порядка (см. возражение Q19 и последующий комментарий). В общем случае, главной характеристикой $\star_{III}$ -утверждения для робота является осознание последним того обстоятельства, что коль скоро он предполагает, что механизмы, обуславливающие 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 (P_1 -

высказывания). В этом рассуждении нет ничего от тех внутренне противоречивых методов рассуждения, к числу которых принадлежит, в частности, парадокс Рассела. Представленные $\star_{III}$ -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, которые «слишком велики» в том или ином смысле)⁶¹.

У робота нет иных причин принимать на веру любое из этих Π_1 -высказываний, кроме как исходя из допущения, что он построен в соответствии с набором правил M , впрочем, для доказательства ему ее вполне хватает. Возникающее впоследствии действительное противоречие не является математическим парадоксом (подобным парадоксу Рассела) – это самое обыкновенное противоречие, связанное с предположением о том, 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 $\star$ -утверждений, полагаемых безошибочными, с целью построения формальной системы 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 Π_1 -высказывание». Мы можем воспользоваться представленными в НРК точными спецификациями машин Тьюринга, положив, что T_m есть не что иное, как « m -тая 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» Π_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 Π_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 Π_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 $\star$ -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 Π_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 Π_1 -высказываний, которые необходимо

⁶¹ Популярное изложение этих вопросов можно найти в [350] [351] и [329].

принимать в рассмотрение в рамках приведенного в §3.20 рассуждения, является конечным, нет никакого явного ограничения на объем доказательств, необходимых роботам для реализации $\star$ -демонстрации истинности всех этих P_1 -высказываний. Даже если ограничить степень сложности принимаемых в рассмотрение P_1 -высказываний самым скромным пределом c , то всё 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 P_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 P_1 -высказывания, то это доказательство неизбежно окажется весьма и весьма сложным и изощренным. Если такое доказательство выдвинет в качестве кандидата на $\star$ -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 $\star$ -статуса). В случае гипотезы Гольдбаха нам неизвестно, является ли это P_1 -высказывание действительно истинным, – а если является, то возможно ли его доказательство в рамках известных и общепринятых методов математического доказательства. Иначе говоря, это P_1 -высказывание может входить в формальную систему Q^* , а может и не входить.

Еще одним «неудобным» P_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 Appel'ем 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 P_1 -высказывания, то степень сложности такого высказывания будет очень небольшой, хотя, наверное, всё же большей, нежели степень сложности P_1 -высказывания, необходимого для выражения гипотезы Гольдбаха. Если бы доказательство Appel'я–Хакена было выдвинуто одним из наших роботов в качестве кандидата на получение $\star$ -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 всё 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 P_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 P_1 -высказываний, которые будут значительно сложнее приведенных здесь, хотя степень их сложности и не превысит величины c . Некоторые из таких P_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 P_1 -высказываний, истинность которого может быть однозначно установлена роботами (посредством демонстрации, достаточно убедительной для присвоения высказыванию $\star$ -статуса и успешного преодоления им всех заграждений, установленных с целью обеспечения безошибочности получаемых роботами результатов), автоматически становится теоремой формальной системы Q^* .

Кроме того, возможны и пограничные случаи, приемлемость или неприемлемость (причем грань между этими состояниями весьма тонка) которых определяется строгостью стандартов, необходимых для получения $\star$ -статуса, или тем, насколько точный характер имеют меры предосторожности, установленные с целью обеспечения безошибочности утверждений, принимаемых в качестве «кирпичей» для построения формальной системы Q^* . Точная формулировка системы Q^* будет различной в зависимости от того, полагаем мы такое P_1 -высказывание безошибочным $\star$ -утверждением либо нет. В обычных обстоятельствах эта разница не имеет большого значения, поскольку различные варианты системы Q^* , обусловленные принятием или

отклонением высказывания P , являются логически эквивалентными. Такая ситуация может возникнуть в случае Π_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 $\star$ -утверждений, которые уже приняты как безошибочные, то возникнет эквивалентная система Q^* , причем вне зависимости от того, принимается высказывание P в качестве ее теоремы или нет. С другой стороны, возможны такие Π_1 -высказывания, которые потребуют для своего доказательства каких-то хитроумных логических процедур, выходящих за рамки любых логических следствий из тех $\star$ -утверждений, которые были приняты как безошибочные ранее, при построении системы Q^* . Обозначим получаемую таким образом формальную систему (до включения в нее высказывания P) через Q_0^* , а систему, образующуюся после присоединения к системе Q_0^* высказывания P , через Q_1^* . Система Q_1^* окажется неэквивалентна системе Q_0^* в том, например, случае, если высказыванием P будет гёделевское предположение $G(Q_0^*)$. 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 Q_0^* (присвоив ему гарантирующий безошибочность $\star$ -статус), коль скоро обоснованность этой системы Q_0^* ими же $\star$ -подтверждена. Таким образом, если они принимают систему Q_0^* , то они должны принять и систему Q_1^* (при условии, что степень сложности высказывания $G(Q_0^*)$ не превышает c – а так оно и будет, если значение c выбрано таким, каким мы выбрали его выше).

Необходимо отметить, что наличие либо отсутствие Π_1 -высказывания P в формальной системе никоим образом не влияет на представленные в §§3.19 и 3.20 рассуждения. Само Π_1 -высказывание $G(Q^*)$ принимается за истинное в любом случае, независимо от того, входит высказывание P в систему Q^* или нет.

Могут найтись и другие способы, с помощью которых роботам удастся «перескочить» через ограничения, налагаемые некоторыми ранее принятыми критериями присвоения $\star$ -статуса Π_1 -высказываниям. В этом нет ничего «парадоксального» – до тех пор, пока роботы не попытаются применить подобное рассуждение к тем самым механизмам M , которые обуславливают их поведение, т.е. к собственно системе 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 кто-то говорил, 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 подлинно интеллектуальная система должна располагать какими-то средствами для

разрыва таких петель, тогда как ни один из существующих компьютеров подобными качествами, вообще говоря, не обладает.⁶² (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. Собственно говоря, 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 Π_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 Π_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 неким «невывислительным 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 заиклилось, либо вообще не приходя ни к какому заключению (по той причине, что теперь заиклился 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 заикнется.

А что если ввести в процесс принятия решения о необходимости «выскакивать из системы» (в случае предположительно заиклившегося 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 связи 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 Π_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

⁶² В.Э.: Что это Пенроуз здесь имеет в виду?!?! (Если понимать буквально, то это настолько нелепо, что даже и не верится, что он мог это сказать! Существуют сколько угодно способов прерывания бесконечных циклов в компьютерах – это элементарно). P.S.: А! Он имел в виду, что есть случаи, когда невозможно установить, надо ли цикл прерывать или не надо. Ну, это, конечно, есть такие случаи. Но и для человека тоже есть такие случаи. А то, что может знать человек об этом, может знать и компьютер.

оказаться значительно меньше, нежели в первом случае. Подобные размышления порождают множество неловких вопросов, ломать копыя 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 P_1 -высказываний, неизбежно оказывается, скажем так, не совсем адекватным.

Если мы намерены научиться однозначно устанавливать истинность любого P_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 всё же это вычисление весьма просто описать (особенно если припомнить, что $65536 = 2^{16}$), причем абсолютно очевидно, что в конечном итоге оно всё равно завершится. Если же мы вознамеримся счесть, что вычисление заиклилось 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, внося в программу несколько предложенных Элькисом упрощающих поправок и, в конечном счете, затратив приблизительно 100 часов компьютерного времени, получил несколько, правда, меньшее (вообще говоря, наименьшее 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]⁶³). Мы можем перефразировать предположение Эйлера и записать его в следующем виде: не имеет решения в положительных целых числах уравнение

$$p^n + q^n + \dots + t^n = u^n,$$

где $p, q, \dots, t$ 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 же 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.)

В 1935 году ученик Литлвуда по фамилии Скъюс показал, что упомянутая точка приходится на число, меньшее $10^{10^{34}}$, однако точное число так и остается неизвестным, хотя оно, конечно же, значительно меньше предела, поставленного Скъюсом. (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 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 интерпретации 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

⁶³ 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 всё 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.

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. Однако 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 обуславливает 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 всё 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 N и последующей попытки вывода теорем в рамках этой системы. Из §2.9 нам известно, что отыскание доказательств всех теорем системы N 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. Тебб – 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.⁶⁴

Более близкую аналогию с описанными в предыдущих разделах процедурами можно усмотреть в предпринимаемых различными исследователями на протяжении последних приблизительно десяти лет попытках разработки «искусственно-интеллектуальных» процедур для реализации математического «понимания».⁶⁵ 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, многие с этим согласятся), из этих попыток пока что ничего, что представляло бы сколько-нибудь реальный математический интерес, не вышло. Мне, несомненно, возразят, что мы еще лишь в начале пути, и наверняка в будущем от них 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 механизмам⁶⁶ (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.⁶⁷ Хотя в строгом доказательстве, возможно, еще и остались какие-то крошечные «лазейки»⁶⁸, вряд ли сквозь них можно протолкнуть что-нибудь существенное. Кто-то очень рассчитывает на лазейку под названием «божественное вмешательство» (посредством

⁶⁴ 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.

⁶⁵ Исторический обзор некоторых таких попыток можно найти у Д. Фридмана [124].

⁶⁶ В.Э.: Итак, ядром пенроузовского «доказательства» была теорема Гёделя. Из нее якобы всё и вытекает. Однако теорему Гёделя мы можем воспринимать всерьез только в том случае, если принят «постулат Кантора» об одинаковости всех «счетных множеств». Для «мира» компьютерных (в том числе мозговых) программ такой постулат не в силе (и, следовательно, не в силе теорема Гёделя, и, далее, не в силе доказательство Пенроуза). Если же «постулат Кантора» принять (и тем самым признать теорему Гёделя), то такая система рассуждений сразу перестает относиться к «миру программ». Тогда рассуждения Пенроуза правильны, но они не стоят и «выеденного яйца», так как они провозглашают то же самое, что было уже принято в самом начале «постулатом Кантора»: что ТАКИЕ рассуждения не могут относиться к «миру программ» или к «вычислительным процедурам». Так что – *idem per idem*. Что принял в начале, то и получил в конце. Порочный круг.

⁶⁷ В.Э.: «Выход» есть: не принимать «постулата Кантора», не признавать, что «счетные множества» все имеют одинаковую мощность, не признавать «диагонального метода», тем самым не признавать «теоремы» Кантора, Гёделя и Тьюринга, оставаться строго в рамках «вычислительных процедур», на их основе объяснить деятельность мозга и построить Искусственный интеллект.

⁶⁸ В.Э.: Не о «лазейках» речь: глобально всё неверно.

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.⁶⁹

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 $\star$ -утверждаемых Π_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 Π_1 -высказываний. Применяв 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 действительно просто ошибка, а не какое-то изначально заложенное в систему заблуждение, обнаружить которое роботам мешает 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 неслучайного характера и именно в области этой «границы 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 $\mathcal{A}$), равно как и невозможности эффективного (но бездумного) вычислительного моделирования всех внешних проявлений деятельности разума (точка зрения $\mathcal{B}$). И всё же, несмотря на убедительность этих аргументов, я подозреваю, что очень многим из нас будет чрезвычайно трудно с ними согласиться. Вместо того, чтобы изучить возможность того, что для понимания феномена интеллекта (что бы за этим словом ни стояло) более подходящей окажется точка зрения $\mathcal{C}$ (или даже $\mathcal{D}$), многие приверженцы научного подхода ограничились одними лишь попытками отыскать слабые места в вышеприведенной аргументации, и всё это исключительно ради поддержания упрямой убежденности в том, что точка зрения $\mathcal{A}$ (в крайнем случае, $\mathcal{B}$) непременно должна в конце концов оказаться истинной.⁷⁰

⁶⁹ В.Э.: Интересно, что Пенроуз ответил бы на мою аргументацию? Но, к сожалению, дискуссии у нас с ним не получится: он уже очень стар (почти 80 лет), а я впредь принципиально не поведу дискуссию ни на каком другом языке, кроме русского (даже по-латышски нет). Поэтому вместо Пенроуза пусть мне ответят (если могут, конечно ☺) его русскоязычные сторонники!

⁷⁰ В.Э.: Точку зрения $\mathcal{C}$ (Веданскую теорию) мы принимаем за истинную не потому, что она «непрерывно должна в конце концов оказаться истинной», а потому, что она удовлетворяет обычным критериям научной точки зрения: она минимизирует все постулаты, не требует никаких новых «сущностей» и физических законов, и в то же время объясняет огромное множество всевозможных явлений от оснований математики до исторической слепоты.

Я не считаю такую реакцию неразумной. Точки зрения $\mathcal{C}$ и $\mathcal{D}$ тоже не свободны от фундаментальных противоречий. Если мы верим, в соответствии с $\mathcal{D}$, 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 всё же необходимы наши физические мозги? Совершенно очевидно, что изменение физического состояния мозга может повлечь 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 сплховать и при описании свойств интеллекта.

Что касается точки зрения $\mathcal{C}$, 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 мощности, скорости и точности. В результате перед нами открылся широкий простор для проведения всё более тесных аналогий между тем, что происходит в недрах современных универсальных компьютеров, и всевозможными проявлениями самой 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 методов⁷¹; 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 $\mathcal{C}$. Таким образом, из двух версий $\mathcal{C}$ 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

⁷¹ 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.

деятельность разума включает в себя какие-то «невыхислительные» процессы), однако вместе с тем они были убеждены в том, что для отыскания этих самых «невыхислительных» 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, обуславливающих 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,⁷² 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»)⁷³.

Хотя окончательных и недвусмысленных математических теорем на этот счет в нашем распоряжении практически нет,⁷⁴ всё же вряд ли кто-либо всерьез сомневается в том, что все существующие физические теории являются по своей природе и в своей основе вычислимыми – 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, всё 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 хитроумное 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, эта 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

⁷² 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.

⁷³ 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].

⁷⁴ См., например, [346], [316], [29] и [328].

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 невывислимые 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 Π_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 всё. 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 будем обсуждать 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 всё 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, обуславливающие 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»⁷⁵). Ничто не мешает нам вообразить, что в результате такой дарвиновской эволюции может развиваться 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, обуславливают функционирование нашего мозга, не могут иметь непосредственного отношения к пониманию деятельности человеческого разума. Они согласны с тем, что поведение человека действительно «невыхислимо», однако полагают, что эта невыхислимость 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 обуславливающих 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.

⁷⁵ Замечательные описания игры Конуэя «Жизнь» можно найти в [137], [311] и [391].

Впрочем, даже согласившись со всем вышеизложенным, можно всё 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 обуславливающих 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,⁷⁶ 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, признают, 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, – просто 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 вкуче с многочисленными последователями).

И всё 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 обуславливают осознание и понимание.

Исследование упомянутых материй, помимо прочего, неизбежно затронет и глубинные вопросы философии математики.⁷⁷ 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

⁷⁶ Подробное описание этих экспериментов приведено в [40].

⁷⁷ В.Э.: Ну безусловно! Нынешняя «философия математики» в корне ошибочна, фундаментально неверна (что и порождает бесконечные проблемы математики). Однако принятие правильной «философии математики» (которую дает Веданская теория) означает для «нынешней математики», точнее, для нынешних математиков катастрофу неимоверных масштабов – таких масштабов, каких еще не было в истории математики. Это означает, что практически весь XX век математики работали вхолостую, что пусты и ошибочны все эти «теории множеств», «теоремы Гёделя» и прочее, и прочее. Это означает, что тысячи профессоров в сотнях университетов мира всю свою жизнь преподавали многим поколениям студентов полную чушь! Катастрофа для математиков, конечно, будет беспрецедентной.

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?⁷⁸

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.

⁷⁸ В.Э.: На все эти вопросы отвечает Веданская теория.

Часть II.

Новая физика, необходимая для понимания разума

В поисках невычислительной физики разума

Глава 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.⁷⁹ 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 всё те же математические законы. И всё 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.⁸⁰ 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,⁸¹ 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,

⁷⁹ В.Э.: Ну, это сказано так... неточно. На самом деле математика отражает реальный мир, а не реальный мир подчиняется математике. Выражение Пенроуза лучше было бы сформулировать так: Вселенная подчиняется законам, которые мы можем описать в математической форме.

⁸⁰ В.Э.: Это как кому! ☺ Чем примитивнее мозговой компьютер, тем легче ему это делать.

⁸¹ См., напр., [81], с. 49.

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 всё 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»⁸²; в то же время энтропия оказывается связана и с другими, более «материальными» физическими величинами посредством вполне точных математических соотношений.⁸³

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 $\mathcal{C}$ (ради чего, собственно, я всё это и затеял), нежели с любой из остальных ($\mathcal{A}$, $\mathcal{B}$ и $\mathcal{D}$, 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 $\mathcal{C}$. 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).⁸⁴ Отправимся же в путь вслед за нашей путеводной звездой $\mathcal{C}$!

§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 всё же попробуем отыскать в дозволенных законами пределах место для той таинственной невычислительной активности, которая каким-то образом оказывается необходимой⁸⁵ 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

⁸² В.Э.: Ну, это по плохой теории информации (в которую мы сейчас не можем вдаваться). А если смотреть более здраво, то та – их «информация» скорее сама придаток к энтропии (и не имеет отношения к той информации, которая обрабатывается компьютерами – включая мозговой).

⁸³ Одно из таких соотношений «первый закон термодинамики»: $dE = dS = pdV$. Буквами E , T , S , p и V здесь обозначены, соответственно, энергия, температура, энтропия, давление и объем газа.

⁸⁴ В.Э.: «Красоту и мощь» аргументации Первой части книги я «не оценю», но последовать за Пенроузом в область квантовой механики и выслушать его предложения мне, тем не менее, очень интересно. Вполне возможно, что там откроется и что-то пригодное для моей философии и Веданской теории.

⁸⁵ В.Э.: Не оказывается.

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 всё 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.⁸⁶ 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 случайные 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.

⁸⁶ В.Э.: По «Копенгагенской интерпретации». Но возможны и другие интерпретации.

§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 *A*) взгляда на природу сознания,⁸⁷ 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.⁸⁸ 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

⁸⁷ См., напр., [81].

⁸⁸ В.Э.: Мозг должен быть организован так, чтобы в нем протекало самопрограммирование.

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 всё 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 слабо – в сравнении с прочими известными физикам силами. Например, сила гравитационного притяжения между электроном и протоном в атоме водорода в

$$28 \cdot 500 \cdot 000 \cdot 000 \cdot 000 \cdot 000 \cdot 000 \cdot 000 \cdot 000 \cdot 000 \cdot 000 \cdot 000 \cdot 000 \cdot 000$$

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, таким последствием, 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.⁸⁹)

Впрочем, как выяснилось позднее, такое предположение лишено оснований. Главная причина заключается в том, что гравитация воздействует на причинные связи между пространственно-временными событиями; никакая другая физическая величина такого воздействия не производит. Можно сказать иначе: гравитация обладает уникальной способностью «наклонять» световые конусы. (Вскоре я поясню, что всё это означает.) Только гравитация может наклонять

⁸⁹ [333]; см. также [265], с. 428.

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-временными 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 всё рассмотрено более подробно.) На рис. 4.1 я изобразил единичный световой конус в пространственно-временных координатах. Ось времени на рисунке направлена снизу вверх, пространство же «откладывается» по горизонтали. Точкой на пространственно-временной диаграмме отображается событие, т.е. некая точка пространства в какой-то определенный момент времени. Событие, таким образом, имеет нулевую временную продолжительность, равно как и нулевую пространственную протяженность. Полный световой конус с центром в точке-событии P представляет пространственно-временную историю сферического светового импульса, который «схлопывается» внутрь P и тут же «выплескивается» обратно, наружу; всё это, разумеется, происходит со скоростью света. Таким образом, световой конус события P образуют все те лучи света, в индивидуальной истории которых событие P происходило.

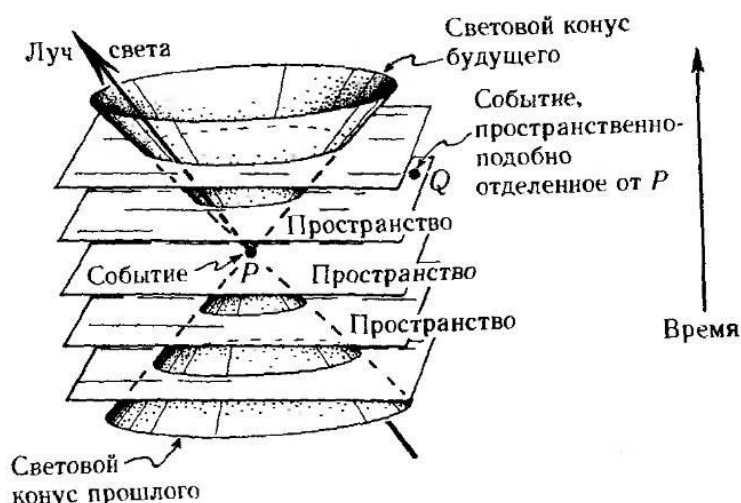


Рис. 4.1. Световой конус события P составляют все те лучи света, которые в пространстве-времени проходят через событие P . Сам конус представляет собой историю вспышки света, схлопывающейся в точку P (световой конус прошлого) и вырывающейся затем наружу (световой конус будущего). События Q и P пространственноподобно разделены (точка Q лежит вне светового конуса P), т.е. событие Q оказывается вне зоны причинного воздействия события P .

Световой конус P состоит из двух частей: светового конуса прошлого⁹⁰ (входящая вспышка) и светового конуса будущего (исходящая вспышка). Согласно теории относительности, причинное воздействие на пространственно-временное событие 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

⁹⁰ На рисунках в НРК изображены только «будущие» части световых конусов.

относительности: 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 большую «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.

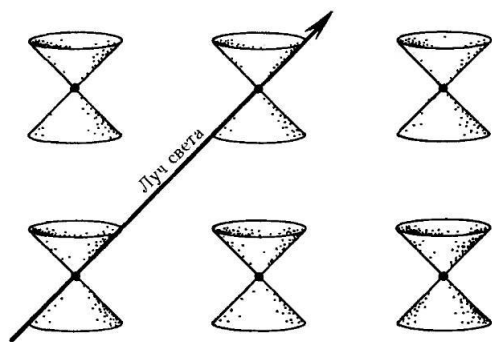


Рис. 4.2. Пространство Минковского: пространство-время в специальной теории относительности. Все световые конусы размещены равномерно и сориентированы в одном направлении.

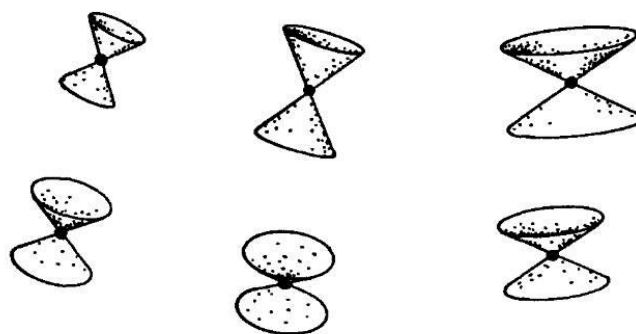


Рис. 4.3. Наклонные световые конусы в общей теории относительности Эйнштейна.

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 световых 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.

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 этаким всепроникающей и повсеместной преломляющей средой, которая оказывает воздействие не только на поведение реального света, но и на поведение всех материальных

частиц и сигналов.⁹¹ 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 способность 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.



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.

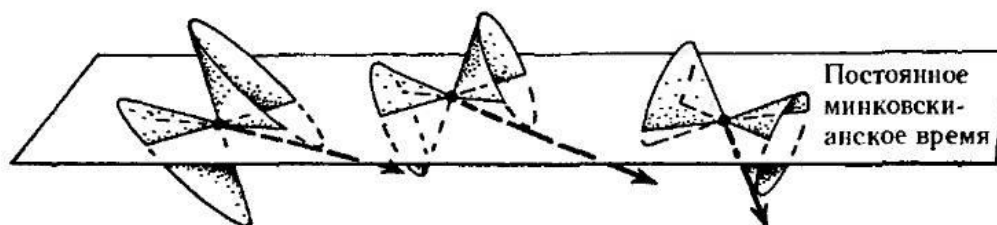


Рис. 4.5. В принципе наклон светового конуса может стать настолько большим, что сигналы смогут распространяться в минковскианское прошлое.

Известны и гораздо более «экстремальные» ситуации, в которых описать таким образом наклон световых конусов и вовсе невозможно, даже если допустить «превышение» абсолютной

⁹¹ Забавно, что сам Ньютон тоже высказывал подобную идею. (См. «Вопросы» 18–22 в третьей книге «Оптики» (1730).)

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).

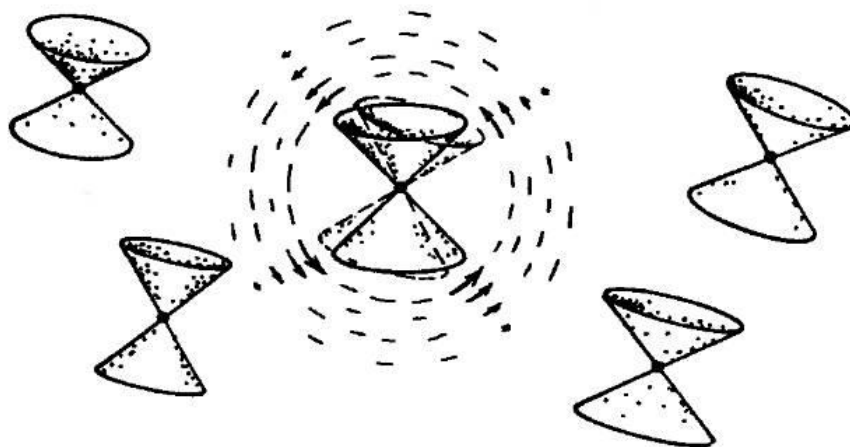


Рис. 4.6. Вообразим пространство-время в виде резинового листа с нанесенными на нем световыми конусами. Каждый отдельный световой конус можно поворачивать (растягивая резину) до тех пор, пока все они не выстроятся в стандартную минковскианскую картину.

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 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 Минковского⁹² (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 «странные» 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

⁹² В.Э.: Видимо, плохой перевод; надо бы: «в пространстве Минковского специальной относительности».

пространственно-временной 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.

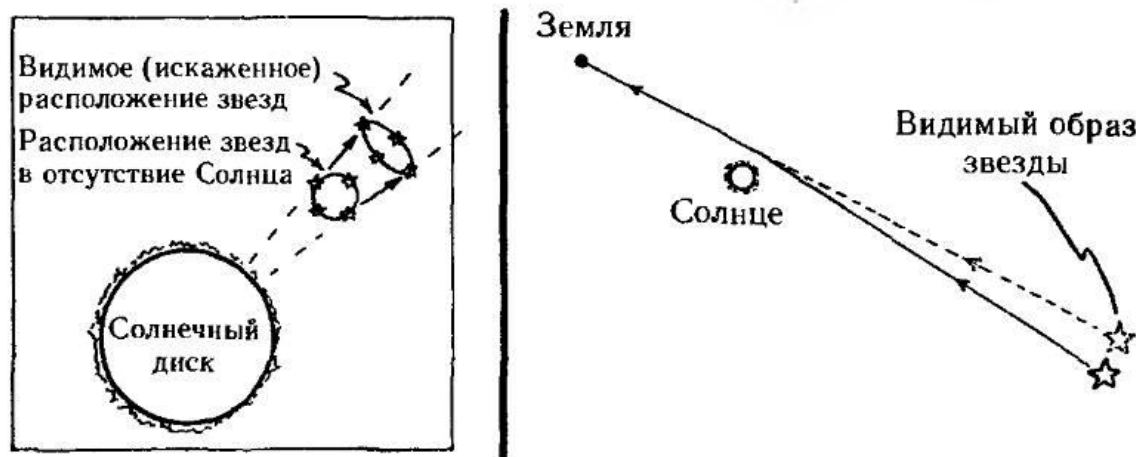


Рис. 4.7. Непосредственно наблюдаемый эффект наклона световых конусов. Пространственно-временное WEYL-искривление проявляется в виде искажения картины звездного неба в результате отклонения световых лучей под воздействием гравитационного поля Солнца. Круг из звезд представляется наблюдателю эллипсом.

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 (вкуче с эффектами временной задержки) можно извлечь весьма ценные сведения о соответствующих расстояниях, массах и т.д. Всё это можно полагать достаточно недвусмысленным свидетельством в пользу того, что феномен наклона световых конусов действительно существует, а также того, что WEYL-эффекты непосредственно измеримы.

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.



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.

Гравитационные взаимодействия наблюдались на примере объектов, значительно больших, нежели песчинки, но всё 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 обуславливает 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 немислимо 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).

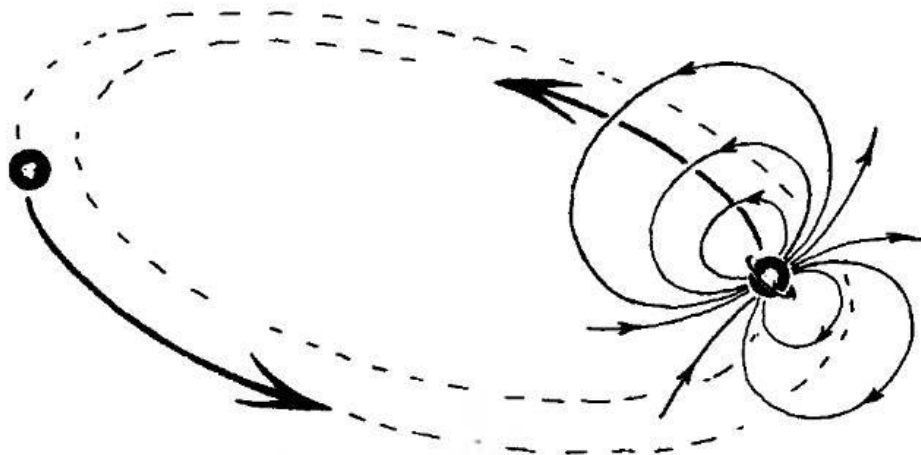


Рис. 4.9. Двойной пульсар PSR 1913+16. Две нейтронные звезды вращаются одна вокруг другой. Одна из звезд является пульсаром; ее магнитное поле чрезвычайно велико и способно захватывать заряженные частицы.

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 принимают 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).

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 вычисле-

ния 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.⁹³ 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.



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.

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 существенно 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 через 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!

⁹³ **В.Э.:** Из всех применений математики меня с самого детства больше всего впечатляли вычисления о звездах (такие, как об этом PSR 1913+16). Там – в невообразимой дали (в данном случае – на расстоянии 30'000 световых лет – это значит, что те радиоволны, которые мы теперь ловим, были отправлены «в эфир» тогда, когда на Земле не существовало еще ни одной цивилизации, и кроманьонцы лишь начинали что-то рисовать на скалах, приручили собаку и стали ловить рыбу) – в невообразимой дали находятся какие-то тела, а мы – люди! – можем вычислить их массы, орбиты, периоды обращения, плотности и т.д.! Потрясающая сила математики! Однако это примечание я сделал не для того, чтобы просто повосхищаться математикой. В связи с ЭТОЙ книгой я хочу напомнить читателю: все эти чудеса сотворяет НАСТОЯЩАЯ математика – та, которая возникает из мозговых программ, работающих с объектами внешнего мира, та, которая видна в четырех факсимильных страницах Предисловия настоящего тома. А то «около-математическое построение», которое возникает от «формализации», «аксиоматизации» и «диагонального метода» – все эти «теоремы Кантора», «теоремы Гёделя», «теоремы Тьюринга» и «трансфинитные числа», все эти Q , Q^* , $Q(M)$, $Q_{\mathcal{M}}(M)$, $Q'_{\mathcal{M}}(M)$, $Q'_{\mathcal{M}}(M, c)$ – бесплодны как пустыня. Ничего ни о звездах, ни о чем другом на их основе не вычислишь и не узнаешь. Вот так!

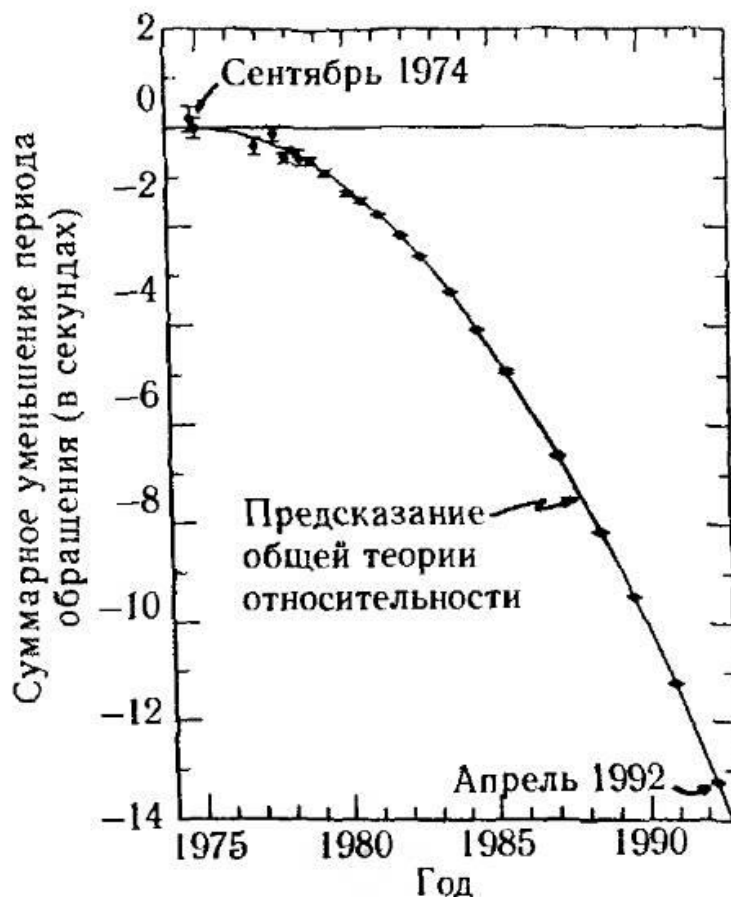


Рис. 4.11. Этот график (любезно предоставленный Дж. Тейлором) демонстрирует точное согласие наблюдаемого (на протяжении 20 лет) уменьшения периода взаимного обращения составляющих пульсар нейтронных звезд с расчетной потерей энергии системой при гравитационном излучении в соответствии с теорией Эйнштейна.

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!

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 – всё 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 всё 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.⁹⁴ 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

⁹⁴ 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].

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 возникновению 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 образец 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 интенсив-

ностью 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 COBE⁹⁵); непрерывная кривая построена в соответствии с формулой Планка, при этом за температуру фонового излучения взято значение 2,735 ($\pm 0,06$) К (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 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,

⁹⁵ Cosmic Background Explorer (англ.) – букв. «Исследователь космического фонового излучения». – Прим. перев.

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 эффекты (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 беспристрастности такого предположения в отношении моделирования многих биологических систем (сюда, пожалуй, можно включить и распространение нервных импульсов) всё 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.

(Продолжение в книге {PENRS3})

Векордия (VEcordia) представляет собой электронный литературный дневник Валдиса Эгле, в котором он цитировал также множество текстов других авторов. Векордия основана 30 июля 2006 года и первоначально состояла из линейно пронумерованных томов, каждый объемом приблизительно 250 страниц в формате A4, но позже главной формой существования издания стали «извлечения». «Извлечение Векордии» – это файл, в котором повторяется текст одного или нескольких участков Векордии без линейной нумерации и без заранее заданного объема. Извлечение обычно воспроизводит какую-нибудь книгу или брошюру Валдиса Эгле или другого автора. В названии файла извлечения первая буква «L» означает, что основной текст книги дан на латышском языке, буква «E», что на английском, буква «R», что на русском, а буква «M», что текст смешанный. Буква «S» означает, что файл является заготовкой, подлежащей еще существенному изменению, а буква «X» обозначает факсимилы. Файлы оригинала дневника Векордия и файлы извлечений из нее Вы **имеете право** копировать, пересылать по электронной почте, помещать на серверы WWW, распечатывать и передавать другим лицам бесплатно в информативных, эстетических или дискуссионных целях. Но, основываясь на латвийские и международные авторские права, **запрещено** любое коммерческое использование их без письменного разрешения автора Дневника, и **запрещена** любая модификация этих файлов. Если в отношении данного текста кроме авторских прав автора настоящего Дневника действуют еще и другие авторские права, то Вы должны соблюдать также и их.

В момент выпуска настоящего тома (обозначенный словом «Версия:» на титульном листе) главными представителями Векордии в Интернете были сайты: для русских книг – <http://vecordija.blogspot.com/>; для латышских книг – <http://vekordija.blogspot.com/>.

Оглавление

VEcordia	1
Извлечение R-PENRS2	1
Роджер Пенроуз	1
ТЕНИ РАЗУМА	1
Предисловие к Третьей главе Пенроуза	2
Роджер Пенроуз. «Тени разума»	4
Глава 3. О невычислимости в математическом мышлении	4
§3.1. Гёдель и Тьюринг	4
§3.2. Способен ли необоснованный алгоритм познаваемым образом моделировать математическое понимание?	7
§3.3. Способен ли познаваемый алгоритм непознаваемым образом моделировать математическое понимание?	9
§3.4. Не действуют ли математики, сами того не осознавая, в соответствии с необоснованным алгоритмом?	14
§3.5. Может ли алгоритм быть непознаваемым?	18
§3.6. Естественный отбор или промысел Господень?	20
§3.7. Алгоритм один или их много?	21
§3.8. Эзотерические математики не от мира сего как результат естественного отбора	22
§3.9. Алгоритмы обучения	25
§3.10. Может ли окружение вносить неалгоритмический внешний фактор?	27
§3.11. Как обучаются роботы?	28
§3.12. Способен ли робот на «твердые математические убеждения»?	30
§3.13. Механизмы математического поведения робота	33
§3.14. Фундаментальное противоречие	35
§3.15. Способы устранения фундаментального противоречия	37
§3.16. Необходимо ли роботу верить в механизмы М?	38
§3.17. Робот ошибается и робот «имеет в виду»?	40
§3.18. Введение случайности: ансамбли всех возможных роботов	41
§3.19. Исключение ошибочных ☆-утверждений	42
§3.20. Возможность ограничиться конечным числом ☆ _m -утверждений	44
§3.21. Окончателен ли приговор?	47
§3.22. Спасет ли вычислительную модель разума хаос?	48
§3.23. <i>Reductio ad absurdum</i> – воображаемый диалог	49
§3.24. Не парадоксальны ли наши рассуждения?	57
§3.25. Сложность в математических доказательствах	59
§3.26. Разрыв вычислительных петель	61
§3.27. Вычислительная математика: процедуры нисходящие или восходящие?	64
§3.28. Заключение	66
Часть II. Новая физика, необходимая для понимания разума	74
Глава 4. Есть ли в классической физике место разуму?	74
§4.1. Разум и физические законы	74
§4.2. Вычислимость и хаос в современной физике	75
§4.3. Сознание: новая физика или «эмергентный феномен»?	77
§4.4. Эйнштейнов наклон	77
§4.5. Вычисления и физика	85
Оглавление	93