

# VEcordia

## Извлечение R-PENRS1

Открыто: 2010.08.10 01:43  
Закрито: 2010.08.21 16:00  
Версия: 2016.12.10 15:47

ISBN 9984-9395-5-3

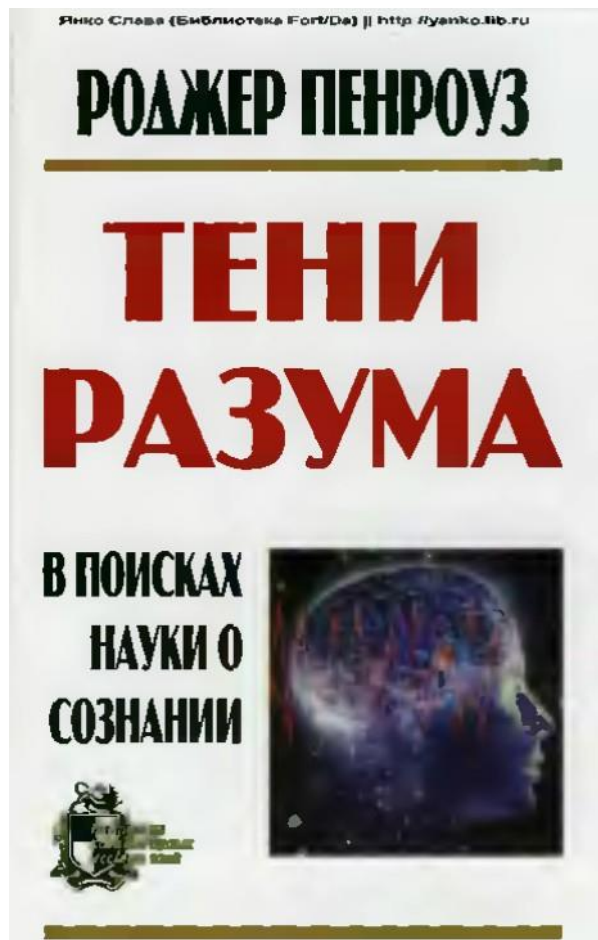
Дневник «VECORDIA»

© Valdis Egle, 2016

ISBN 5-93972-457-4

Роджер Пенроуз. «Тени Разума», том I

© Роджер Пенроуз, 1994



Обложка русского издания

Роджер Пенроуз

# ТЕНИ РАЗУМА

Часть I

Невычислимость сознательного мышления

(главы 1 – 2)

С комментариями Валдиса Эгле

Impositum

Grīziņkalns 2016

Talis hominis fuit oratio,  
qualis vita

## Предисловие в Векордии

Как математика, связанного с передним краем физических исследований, я Роджера Пенроуза знал уже раньше, но как автор концепций, противоречащих Веданской теории, он появился на моем «горизонте» в 1999 году, когда вышла моя книга «*Tur tālumā, kur ziemas perazīst*» с изложением по-латышски основ Веданской теории. Тогда профессор Тамберг и доктор Балодис указали мне на книгу Пенроуза «*The Emperor's New Mind*»; один ее экземпляр (не типографский, а переплетенная ксерокопия) нашелся в Академической библиотеке Латвии, и там я ее прочел (по-английски), сделав также себе копии наиболее важных глав.

На основные аргументы Пенроуза, на которые указывали Тамберг и Балодис, я им тогда ответил; я знал также, что Пенроуз написал еще и вторую книгу, но она была мне недоступна: ее не было ни в латвийских библиотеках, ни в Интернете (где я периодически ее искал).

Включить обе книги Пенроуза в Векордию и проанализировать их с точки зрения Веданской теории всегда было моей мечтой, но долгое время неосуществимой из-за недоступности для меня этих книг. И вот, вчера, 9 августа 2010 года, очередной раз (после, может быть, 2 или 3-летнего перерыва) запросив снова их поиск в Интернете, я наконец обнаружил их тексты на русском языке в виде *Djvu* файлов. Этот формат мало приспособлен для переноса в Векордию, особенно если текст столь сложен (в смысле всяких там математических значков), как у Пенроуза. Но, тем не менее, начало этому мероприятию я положил.

Я спланировал четыре тома Роджера Пенроуза в Векордии:

R-PENRO1	«Новый Разум Короля», часть I
R-PENRO2	«Новый Разум Короля», часть II
R-PENRS1	«Тени Разума», часть I
R-PENRS2	«Тени Разума», часть II

Но первым я, разумеется, собираюсь сделать третий из этих томов, потому, что он имеет столь очаровательное название: «Почему для понимания разума необходима новая физика? Невычислимость сознательного мышления».

На этот материал (Первую часть «Теней Разума») разные авторы ссылаются так часто, как это бывает редко с какой книгой, – ссылаются все, кто действительно поверил (а большинство людей поверили!), что человеческий интеллект и вправду невозможно реализовать на компьютерах. Поэтому именно этот материал (PENRS1 по названному проекту) и представляет для меня главнейшую ценность.

Сейчас, когда я пишу это Предисловие, я не читал еще «Теней Разума» и не знаю, как Пенроуз собирается доказывать, что невозможно написать ту программу, которую я знаю как написать. Правда, я имею некоторое представление об этом по Первой книге Пенроуза, но, говорят, во Второй книге он еще хлеще и уж точно неопровержимо всё это доказывает. И вот – я теперь горю желанием поскорее узнать: как же он это делает?!

Что же, читатель, Вы будете свидетелем такого произошедшего в действительности процесса: я буду постепенно перекачивать сюда, в этот том, текст Пенроуза и одновременно его (впервые) читать, обдумывать и по ходу дела высказывать свои комментарии<sup>1</sup> (как обычно, в сносках, снабженных инициалами «В.Э.:», а если вдруг сказать понадобится так много, что в сноску это не влезет, – тогда прямо в виде вставок в текст Пенроуза).

Как читатель, наверное, чувствует по моему тону, я не верю, что Пенроуз действительно докажет «Невычислимость сознательного мышления». (Невозможно доказать то, что противоречит истине). Ситуация ведь и вправду забавная: я, бывший (теперь на пенсии) старший научный сотрудник Института электроники и вычислительной техники АН ЛатвССР, профессиональный программист и проектировщик больших программных систем, создавший множество компьютерных программ, в том числе операционную систему Диспос (DISPOS), которая эксплуатировалась

---

<sup>1</sup> Ведь это же дневник, а не учебник!

(не мной, а персоналом) в течение 15 лет, – я спроектировал еще одну систему вдобавок ко всем предыдущим – на этот раз операционную систему под названием Витос (VITOS) – и представляю, как ее реализовать, представляю примерно в такой же мере, в какой представлял систему Диспос перед началом ее реализации...

И тут вдруг приходит Роджер Пенроуз и говорит, что Витос невозможно написать. По каким-то там его теоретическим соображениям...

Естественно, что я ему не верю. (А кто такой Пенроуз? Умеет ли он вообще проектировать операционные системы? Создал ли он свою операционную систему, и работала ли она 15 лет с чужим персоналом?... По-моему, он только разъезжал по миру и лекции читал, и никогда ничего на компьютерах не программировал – во всяком случае большое).

Я не верю, что он что-то докажет МНЕ – когда я знаю, как строить Витос, – но мне очень интересно узнать: в чем же он собьется, в чем загвоздка: почему он приходит к таким (нелепым) выводам?

Это и составляет интригу настоящего исследования.

Валдис Эгле

10 августа 2010 года

## Роджер Пенроуз. «Тени разума»

### В поисках науки о сознании

Пенроуз Р.

Тени разума: в поисках науки о сознании.

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

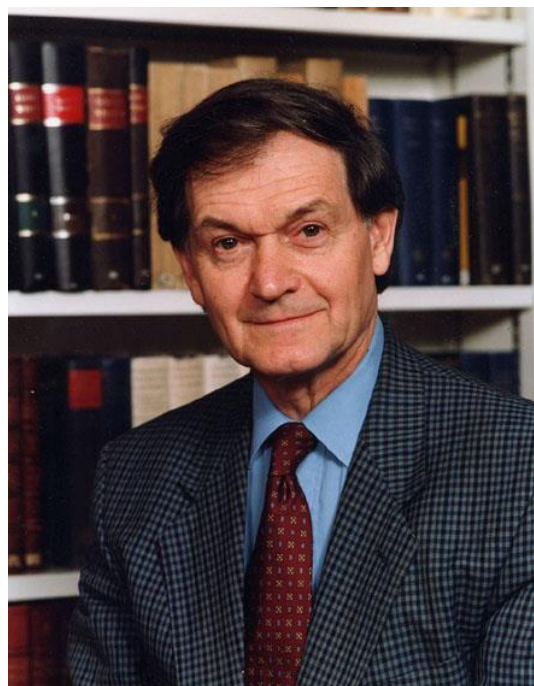
Москва – Ижевск: Институт компьютерных исследований, 2005. 688 с.

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

Для широкого круга читателей, интересующихся наукой.

© Roger Penrose 1994

© Перевод на русский язык: Институт компьютерных исследований, 2005



Роджер Пенроуз – р. 8 августа 1931 г.

This translation of *Shadows of the Mind* originally published in English in 1994 is published by arrangement with Oxford University Press.

Данный перевод книги «Тени разума», оригинальное издание которой было выпущено в 1994 году на английском языке, публикуется с разрешения Oxford University Press.

## Предисловие

Эту книгу можно считать, в некотором смысле, продолжением «Нового разума короля»<sup>2</sup> (далее НРК). То есть я и в самом деле намерен продолжить развитие темы, начатой в НРК, однако излагаемый здесь материал можно рассматривать и совершенно независимо от предыдущей книги. Отчасти необходимость в повторном обращении к предмету первоначально возникла из желания дать как можно более обстоятельные ответы на множество вопросов и критических замечаний, которыми самые разные люди отреагировали на рассуждения и доказательства, представленные в НРК. Тем не менее, тема новой книги представляет собой совершенно самостоятельное исследование, а предлагаемые здесь идеи отнюдь не ограничиваются рамками, установленными в НРК. Одну из главных тем НРК составило мое убеждение в том, что, используя сознание, мы способны выполнять действия, не имеющие ничего общего с какими бы то ни было вычислительными процессами. Однако в НРК эта идея была представлена лишь как осторожная гипотеза; имелась также некоторая неопределенность относительно того, какие именно типы процедур следует включать в категорию «вычислительных процессов». На страницах же этой книги, как мне представляется, читатель найдет гораздо более последовательное и строгое обоснование приведенного выше общего утверждения, причем представляемое обоснование оказывается применимо ко всем типам вычислительных процессов, какие только можно вообразить. Кроме того, здесь имеется и существенно более правдоподобное (нежели это было возможно во времена НРК) предположение относительно механизма церебральной активности, посредством которого наше управляемое сознанием поведение может основываться на какой-либо физической активности невычислительного характера.

Упомянутое обоснование проводится по двум различным направлениям. Одно из них по сути своей негативно; здесь я решительно выступаю против широко распространенного мнения, согласно которому нашу сознательную мыслительную деятельность – во всех ее разнообразных проявлениях – можно, в принципе, адекватно описать в рамках тех или иных вычислительных моделей.<sup>3</sup> Другое направление моих рассуждений можно счесть позитивным в том смысле, что оно предполагает подлинный поиск (разумеется, в рамках необходимости придерживаться строгих и неопровержимых научных фактов) инструментов, позволяющих описываемому в научных терминах мозгу применять для осуществления требуемой невычислительной деятельности тонкие и по большей части нам пока не известные физические принципы.

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

---

<sup>2</sup> The Emperor's New Mind. (Не так давно книга была переведена на русский язык: Пенроуз Р. *Новый ум короля*, М.: Едиториал УРСС. 2003.) – Прим. перев. **В.Э.**: См. {PENRO1}, {PENRO2}, {PENRO3}, {PENRO4}, {PENRO5}.

<sup>3</sup> **В.Э.**: Вообще термин «вычислительный» очень неудачен; неудачен уже и английский «computing», но русский еще в большей мере; он предполагает какие-то ВЫЧИСЛЕНИЯ (т.е. работу с числами – что ли?). Точный термин – и он постоянно используется в Веданской теории – это: обработка информации. Человеческая умственная деятельность представляет собой обработку информации (а не вычисления!); возможно, это уже первая развилка, где Пенроуз уходит не по той тропинке.

также подробные доказательства невозможности достижения системами восходящего (равно как и нисходящего) типа подлинной разумности. В заключение делается вывод о том, что сознательное мышление и в самом деле должно включать в себя процессы, которые с помощью одних лишь вычислительных методов невозможно даже адекватно смоделировать; еще менее способны вычисления, взятые сами по себе, обусловить какое бы то ни было сознательное ощущение или желание. Иными словами, разум, по всей видимости, представляет собой такую сущность, которую никоим образом невозможно описать посредством каких бы то ни было вычислений.<sup>4</sup>

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

Я полагаю, что именно на этом уровне в действие должна вступать физическая невычислимость – условие, необходимое для объяснения невычислимости деятельности сознания. В соответствии с этим предположением я должен потребовать, чтобы уровень, на котором становится значимой упомянутая физическая невычислимость, играл особую роль и в функционировании мозга.<sup>5</sup> Именно в этом пункте мои нынешние предположения наиболее существенно расходятся с теми, что были высказаны в НРК. Я утверждаю, что, хотя сигналы нейронов и могут вести себя как детерминированные в классическом смысле события, управление синаптическими связями между нейронами происходит на более глубоком уровне, т.е. там, где можно ожидать наличия существенной физической активности на границе между квантовыми и классическими процессами. Выдвигаемые мною специфические предположения требуют возникновения внутри микроканалцев цитоскелета нейронов макроскопического квантовокогерентного поведения (в точном соответствии с предположениями Фрёлиха). Иначе говоря, я полагаю, что упомянутая квантовая активность должна быть неким невычислимым образом связана с поддающимся вычислению процессом, который, как утверждают Хамерофф и его коллеги, имеет место внутри этих самых микроканалцев.

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

---

<sup>4</sup> В.Э.: Интересно, с какого же момента филогенеза появляется эта «невычислимость»? Присуща ли она только человеку? А обезьянам? А собакам? А курам? И оса-сфекс, описанная Жаном Фабром, тоже имеет эту «невычислимость»? Посмотрим, даст ли Пенроуз ответы на эти вопросы.

<sup>5</sup> В.Э.: Отметим, что Пенроуз открыто вводит новый постулат по сравнению с «традиционной наукой». Разумеется, такие постулаты вводить можно – но только в том случае, если без них нельзя обойтись. Если же всё можно объяснить и без них (а в данном случае – можно), то начинает действовать Лезвие Оккама. Кроме того, Пенроуз опирается не на существующие сейчас законы квантовой механики, а на те, которые еще только должны быть открыты! (Такое положение многократно рассматривалось в бюллетенях «В Защиту Науки» как типичный признак лженаук).

действительности управляющие этим самым «материалом», из которого состоит наш физический мир.

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

Р.П.

Оксфорд, апрель 1994<sup>6</sup>

## **Благодарности**

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

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

Есть, наконец, еще три человека, которые заслуживают особого упоминания. Невозможно переоценить бескорыстную помощь и поддержку, которую оказал мне Энгус Макинтайр, проверив мои рассуждения относительно математической логики в главах 2 и 3 и предоставив мне множество полезной литературы. Выражаю ему свою глубочайшую благодарность. Стюарт Хамерофф рассказал мне о цитоскелете и его микроканальцах; два года назад я и не подозревал о существовании подобных структур! Я очень ему благодарен за эту бесценную информацию, а также за помощь, которую он оказал мне, проверив большую часть материала главы 7. Я навеки у него в долгу за то, что он открыл моим глазам чудеса нового мира. Он, равно как и все остальные, кого я здесь благодарю, конечно же, ни в коей мере не ответственен за те ошибки, совсем избавиться от которых нам так и не удалось. Особо признателен я своей любимой Ванессе<sup>7</sup> по нескольким причинам: за то, что она объяснила мне, почему отдельные части этой книги нужно переписать; за помощь с литературой, что просто спасло меня, а также за ее любовь,

<sup>6</sup> В.Э.: Пенроузу было 62 года, когда он подписывал это Предисловие; сегодня (2010.08.10) ему 79 лет и 2 дня.

<sup>7</sup> В.Э.: Ванесса Томас – вторая жена Пенроуза, с которой он имеет одного ребенка; с первой женой Джоан Изабелой Ведж он имеет трех сыновей.

терпение и понимание, особенно если учесть, что я постоянно недооцениваю то количество времени, которое отнимает у меня написание книги! Ах, да, чуть не забыл: еще я благодарен ей за то – она, кстати, об этом ничего не знала, – что она отчасти послужила моделью для вымышленного образа Джессики, героини придуманной мною истории. Мне очень жаль, что я совсем не знал Ванессу, когда ей было столько же лет, сколько Джессике!

## Читателю

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

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

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

Соответственно, в этой книге я избрал политику повсеместного употребления в отношении той или иной абстрактной личности местоимений «он», «ему» или «его»<sup>8</sup>. Из этого ни в коем случае не следует делать вывода о половой принадлежности упомянутой личности. Эту личность не нужно считать ни мужчиной, ни женщиной. Как правило, индивидуум, которого я называю «он», обладает сознанием и чувствами, а потому называть его «оно»<sup>9</sup>, по-моему, не годится. Я искренне надеюсь, что ни одна из моих читательниц не усмотрит личного оскорбления в том,

---

<sup>8</sup> В.Э.: Столько говорил, говорил в угоду феминисткам, а в конце всё равно раскрыл своё нутро мужского шовиниста! ☺

<sup>9</sup> В оригинале «it» – местоимение третьего лица единственного числа, которым в английском языке называют животных и неодушевленные предметы, независимо от их пола и/или рода. – *Прим. перев.*

что, говоря в §5.3, §5.18 и §7.12 о своем трехглазом коллеге с α-Центавры (абстрактном, разумеется), я использую местоимение «он» и что это же местоимение я употребляю в отношении совершенно безличных индивидуумов в §1.15, §4.4, §6.5, §6.6 и §7.10. Я также надеюсь, что ни один из моих читателей не будет обижен тем, что я использую местоимение «она» в отношении умной паучихи из §7.7 и преданной чуткой слоники из §8.6 (хотя бы по той простой причине, что в этом случае из контекста очевидно, что обе они действительно относятся к женскому полу), а также в отношении демонстрирующей сложное поведение парамеции из §7.4 (которую я отношу к «женскому» роду по не совсем удовлетворительной причине ее прямой способности к воспроизведению себе подобных), ну и самой матушки-Природы в §7.7.

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

$$22 / 17 \times n,$$

где  $n$  – номер страницы книги в твердой обложке, приводимый здесь в качестве ссылки.<sup>10</sup>

## Пролог

Джессика всегда немного нервничала, входя в эту часть пещеры.

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

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

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

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

Джессику это несколько не успокоило.

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

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

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

Джессика ненавидела отца, когда у него бывало такое настроение. Хотя – нет: она всегда любила его, любила больше всего и больше всех, но всегда хотела, чтобы он никогда не становился таким, как сейчас. Она знала, что это настроение каким-то образом связано с тем, что он ученый, но до сих пор не понимала каким именно. Она даже надеялась, что сама когда-нибудь сможет стать ученым, хотя уж она-то позаботится о том, чтобы не впасть в такое состояние духа.

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

---

<sup>10</sup> В.Э.: В русском издании этой книги сохранены ссылки на номера страниц английской НРК, что для нас, разумеется, совершенно бесполезно. В своей электронной публикации я (переходами гипертекста) попытался указать те места НРК, на которые Пенроуз мог бы ссылаться. Но, конечно, мои догадки могут быть неточными.

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

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

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

Джессика вдруг вспомнила, почему так сильно любит папу.

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

Отец немного удивленно посмотрел на нее, но промолчал.

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

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

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

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

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

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

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

– Отлично, папа, и какая у нас теория?

– Не так быстро... минутку... вот: Земля вертится вокруг Солнца!

– Также мне новая теория!

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

– А сложно было бы их убедить?

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

– Ух ты, а это на самом деле правда?

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

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

Отец Джессики едва заметно покраснел:

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

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

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

– Знаешь что? Я согласна с тобой, папа. Этот валун запросто провисит здесь еще двадцать три века и даже дольше!<sup>11</sup>

---

<sup>11</sup> В.Э.: Такое переложение платоновской мысли о тенях в пещере... Видимо, автор будет на этот пример ссылаться, когда речь пойдет о «мире идей»...

## Часть I.

### Почему для понимания разума необходима новая физика?

#### Невычислимость сознательного мышления

## Глава 1. Сознание и вычисление

### §1.1. Разум и наука

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

Кое-кто склонен верить, что мы действительно способны приблизиться к научному пониманию сознания, что в этом феномене вообще нет ничего загадочного, а всеми существенными его ингредиентами мы уже располагаем.<sup>12</sup> Они утверждают, что в настоящий момент наше понимание мыслительных процессов человека ограничено лишь крайней сложностью и изощренностью организацией человеческого мозга<sup>13</sup>; разумеется, эту сложность и изощренность недооценивать ни в коем случае не следует, однако принципиальных препятствий для выхода за рамки современной научной картины нет.<sup>14</sup> На противоположном конце шкалы расположились те, кто считает, что мы не можем даже надеяться на адекватное применение холодных вычислительных методов бесчувственной науки к тому, что связано с разумом, духом да и самой тайной сознания человека.

В этой книге я попытаюсь обратиться к вопросу сознания с научных позиций. При этом, однако, я твердо убежден (и основано это убеждение на строго научной аргументации) в том, что в современной научной картине мира отсутствует один очень важный ингредиент. Этот недостающий ингредиент совершенно необходим, если мы намерены хоть сколько-нибудь успешно уместить центральные проблемы мыслительных процессов человека в рамки логически последовательного научного мировоззрения.<sup>15</sup> Я утверждаю, что сам по себе этот ингредиент не находится за пределами, доступными науке, хотя в данном случае нам, несомненно, придется в некоторой степени расширить наш научный кругозор. Во второй части книги я попытаюсь

---

<sup>12</sup> В.Э.: Веданская теория располагает; остальной мир, видимо, не располагает, если судить по тому бешеному сопротивлению и отрицанию, которые она встречает на протяжении более тридцати лет своего существования. Если бы меня попросили указать тот «ингредиент», которого не достает «остальному миру» по сравнению с Веданской теорией и из-за которого мир не может понять всё так, как оно есть, то я бы ответил так: пожалуй, это самопрограммирование. Они не могут понять, как одни программы могут создавать другие программы совершенно без участия чего-то постороннего. (А не могут понять в основном потому, что у них нет точного и адекватного представления о программах вообще).

<sup>13</sup> В.Э.: Мозг сложен, но сложен главным образом громадным количеством элементов. Принципы же, на которых основывается мозговая «операционная система» на самом деле очень просты. (Я на этот раз поставил слова «операционная система» в кавычки, но дальше в настоящей книге я это делать уже не буду: мы говорим действительно об операционной системе определенного вида). Принципы, на которых основывается работа мозга, столь же просты, как и принципы, на которых основывается работа тела, например: мышцы опираются на кости; сердце разгоняет кровь, кровь переносит кислород... и т.д.

<sup>14</sup> В.Э.: Разумеется, нет.

<sup>15</sup> В.Э.: Верно – одного ингредиента не хватает, и это – самопрограммирование мозговых программ. Люди этого явления не понимают. Как только они поймут мозговое самопрограммирование, так всё станет на свои места.

указать читателю конкретное направление, следуя которому он непременно придет как раз к такому расширению современной картины физической вселенной. Это направление связано с серьезным изменением самых основных из наших физических законов,<sup>16</sup> причем я весьма детально опишу необходимую природу этого изменения и возможности его применения к биологии нашего мозга. Даже обладая нынешним ограниченным пониманием природы этого недостающего ингредиента, мы вполне способны указать области, отмеченные его несомненным влиянием, и определить, каким именно образом он вносит крайне существенный вклад в то, что лежит в основе осознаваемых нами ощущений и действий.

Разумеется, некоторые из приводимых мной аргументов окажутся не совсем просты, однако я постарался сделать свое изложение максимально ясным и везде, где только возможно, использовал лишь элементарные понятия. Кое-где в книге всё же встречаются некоторые сугубо математические тонкости, но только тогда, когда они действительно необходимы или каким-то образом способствуют достижению более высокой степени ясности рассуждения. С некоторых пор я уже не жду, что смогу с помощью аргументов, подобных приводимым ниже, убедить в своей правоте всех и каждого, однако хотелось бы отметить, что эти аргументы всё же заслуживают внимательного и беспристрастного рассмотрения<sup>17</sup> – хотя бы потому, что они создают прецедент, пренебрегать которым нельзя.

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

### §1.2. Спасут ли роботы этот безумный мир?

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

---

<sup>16</sup> В.Э.: Вот так, мистер Пенроуз! Чтобы ввести Ваш ингредиент, требуется изменить законы природы. А чтобы ввести мой ингредиент, не требуется менять НИЧЕГО ни в каких законах. (Ну, и чья же система постулатов проще?) Для моего ингредиента нужно просто ПОНЯТЬ (понять то, что мне самому кажется довольно простым, но на самом деле, видимо – если судить по реальному опыту, – требует довольно высокой программистской квалификации).

<sup>17</sup> В.Э.: Ну уж это я вам обещаю, мистер Пенроуз!

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

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

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

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

И всё же не могут ли компьютеры достичь в конечном итоге чего-то большего? Многие специалисты заявляют, что компьютеры обладают потенциалом, достаточным – по крайней мере, принципиально – для формирования искусственного интеллекта, который со временем превзойдет наш собственный.<sup>18</sup> По утверждению этих специалистов, как только управляемые

<sup>18</sup> См., в частности, [162], [263], [267]. В.Э.: Список литературы в {PENRS4}.

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

Когда же нам следует ожидать наступления сего счастливого момента? По данному вопросу у упомянутых специалистов нет единого мнения. Одни говорят о многих столетиях, другие заявляют, будто эквивалентность компьютера человеку будет достигнута всего через несколько десятилетий.<sup>19</sup> Последние обычно указывают на очень быстрый «экспоненциальный» рост мощности компьютеров и основывают свои оценки на сравнении скорости и точности транзисторов с относительной медлительностью и «небрежностью» нейронов. И правда, скорость работы электронных схем уже более чем в миллион раз превышает скорость возбуждения нейронов в мозге (порядка  $10^9$  операций в секунду для транзисторов и лишь  $10^3$  для нейронов)<sup>20</sup>, при этом электронные схемы демонстрируют высокую точность синхронизации и обработки инструкций, что ни в коей мере не свойственно нейронам. Более того, конструкции «принципиальных схем» мозга присуща высокая степень случайности, что, на первый взгляд, представляется весьма серьезным недостатком по сравнению с продуманной и точной организацией электронных печатных плат.

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

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

---

<sup>19</sup> Моравак [267] основывает свои доводы в пользу такого срока на том, какая, по его мнению, часть коры головного мозга успешно реализована в виде модели (речь, в основном, идет о нейронах, расположенных в сетчатке), и на оценке темпов развития компьютерной технологии в ближайшем будущем. Любопытно, что к началу 1994 года он своего мнения не изменил; см. [268].

<sup>20</sup> Микросхема *Intel Pentium* содержит более трех миллионов транзисторов на «кремниевой пластине» размером с ноготь большого пальца, причем каждый из этих транзисторов способен на 113 миллионов полных циклов в секунду.

Таким образом, если поверить самым смелым заявлениям наиболее отъявленных провозвестников искусственного интеллекта и допустить, что компьютеры и управляемые ими роботы в конечном счете – и даже, вероятно, довольно скоро – во всем превзойдут человека, то получается, что компьютеры способны стать чем-то неизмеримо большим, чем просто помощниками нашего интеллекта. Они, в сущности, разовьют свой собственный колоссальный интеллект.<sup>21</sup> А мы сможем обращаться к этому высшему интеллекту за советом и поддержкой во всех своих заботах – и наконец-то появится возможность исправить всё то зло, что мы принесли в этот мир!

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

### §1.3. Вычисление и сознательное мышление

В чем же здесь загвоздка? Неужели всё дело лишь в вычислительных способностях, в скорости и точности работы, в объеме памяти или, быть может, в конкретном способе «связи» отдельных структурных элементов<sup>22</sup>? С другой стороны, не может ли наш мозг выполнять какие-то действия, которые вообще невозможно описать через вычисление<sup>23</sup>? Каким образом можно поместить в такую вычислительную картину нашу способность к осмысленному осознанию – счастья, боли, любви, какого-либо эстетического переживания, желания, понимания и т.п.<sup>24</sup>? Будут ли компьютеры будущего действительно обладать разумом<sup>25</sup>? Влияет ли обладание сознательным разумом на поведение индивида, и если влияет, то как именно? Имеет ли вообще смысл говорить о таких вещах на языке научных терминов; иными словами, обладает ли наука достаточной компетентностью для того, чтобы рассматривать вопросы, относящиеся к сознанию человека?

Мне кажется, что можно говорить, как минимум, о четырех различных точках зрения<sup>26</sup> – или даже крайностях, – которых разумный индивид может придерживаться в отношении данного вопроса:

---

<sup>21</sup> В.Э.: Компьютеры не могут начать это делать самопроизвольно (как это иногда изображали в книгах и фильмах фантастики). Несмотря ни на какие вычислительные мощности, в компьютерах не появится самопрограммирование до тех пор, пока кто-нибудь из людей его не запустит. Но если его запускать и поддерживать, то, конечно, компьютеры могут превзойти людей и в конце концов (при определенных условиях) обойтись и без них.

<sup>22</sup> В.Э.: Нет – ничего из названного. Для того, чтобы программа извлекала квадратный корень, надо написать ее такой, чтобы она делала именно это. И, чтобы пошло самопрограммирование (и далее: сознание, интеллект и т.д.), надо начальные программы написать такими, чтобы они делали именно это – а не что-нибудь другое.

<sup>23</sup> В.Э.: Точнее: через обработку информации... В принципе можно такое предположить, но это будет новым постулатом, постулирующим наличие таких действий мозга. В то же время все действия мозга легко объясняются через обработку информации, и в таком новом постулате нет необходимости.

<sup>24</sup> В.Э.: Запросто! Но это потребует много места – поэтому: не здесь!

<sup>25</sup> В.Э.: Это зависит от того, какие программы на них запущены. Один и тот же компьютер (достаточной мощности) может обладать и не обладать разумом в зависимости от того, какая операционная система там работает. Если запустят мой Витос – будет обладать, если запустят ОС/360 – не будет обладать. Разум – это характеристика софтвера, а не хардвера.

<sup>26</sup> Эти четыре точки зрения были подробно описаны, например, в [215], с. 252 (следует, впрочем, отметить, что условие, называемое автором статьи «тезисом Черча–Тьюринга», является, по своей сути, скорее «тезисом Тьюринга» (в том смысле, в каком я употребляю этот термин в §1.6), нежели «тезисом Черча»).

*А.* Всякое мышление есть вычисление; в частности, ощущение осмысленного осознания есть не что иное, как результат выполнения соответствующего вычисления.<sup>27</sup>

*В.* Осознание представляет собой характерное проявление физической активности мозга; хотя любую физическую активность можно моделировать посредством той или иной совокупности вычислений, численное моделирование как таковое не способно вызвать осознание.

*С.* Осознание является результатом соответствующей физической активности мозга, однако эту физическую активность невозможно должным образом смоделировать вычислительными средствами.

*Д.* Осознание невозможно объяснить в физических, математических и вообще научных терминах.

Точка зрения *Д*, полностью отрицающая взгляды физикалистов и рассматривающая разум как нечто абсолютно неподвластное языку науки, свойственна мистикам; и, по крайней мере, в какой-то степени, такое мировоззрение, видимо, сродни религиозной доктрине. Лично я считаю, что связанные с разумом вопросы, пусть даже и не объясняемые должным образом в рамках современного научного понимания, не следует рассматривать как нечто, чего науке никогда не постичь. Пусть на данный момент наука и не способна сказать в отношении этих вопросов своего веского слова, со временем ее возможности неминуемо расширятся настолько, что в ней найдется место и для таких вопросов, причем не исключено, что в процессе такого расширения изменятся и сами ее методы. Отбрасывая мистицизм с его отрицанием научных критериев в пользу научного познания, я всё же убежден, что и в рамках усовершенствованной науки вообще и математики в частности найдется немало загадок, среди которых не последнее место займет тайна разума. К некоторым из этих идей я еще вернусь в следующих главах книги, сейчас же достаточно будет сказать, что согласиться с точкой зрения *Д* я никак не могу, поскольку твердо намерен двигаться вперед, следуя пути, проложенному наукой. Если мой читатель питает сильное убеждение, что истинным является именно пункт *Д*, в той или иной его форме, я попрошу его потерпеть еще немного и посмотреть, сколько нам удастся пройти вместе по дороге науки, – и попытаться при этом понять, куда, по моему убеждению, эта дорога в конечном счете нас приведет.

Теперь обратимся к противоположной крайности: к точке зрения *А*. Эту точку зрения разделяют сторонники так называемого сильного, или жесткого, искусственного интеллекта (ИИ), иногда для обозначения такой позиции употребляется также термин функционализм,<sup>28</sup> хотя некоторые распространяют термин «функционализм» еще и на определенные варианты пункта *С*. Одни считают *А* единственно возможной точкой зрения, которую допускает сугубо научное отношение. Другие воспринимают *А* как нелепость, которая вряд ли стоит сколь-нибудь серьезного внимания. Существует, несомненно, множество различных вариантов позиции *А*. (Длинный список альтернативных версий вычислительной точки зрения приводится в [344].) Некоторые из них отличаются лишь различным пониманием того, что следует считать «вычислением» или «выполнением вычисления». Есть и такие приверженцы *А*, которые вообще не считают себя «сторонниками сильного ИИ», поскольку придерживаются принципиально иного взгляда на интерпретацию термина «вычисление», нежели та, что предлагается в традиционном понятии ИИ (см. [112]). Я рассмотрю эти вопросы подробнее в §1.4. Пока же достаточно будет понимать под «вычислением» такую операцию, какую способны выполнять обычные универсальные компьютеры. Другие сторонники позиции *А* могут расходиться в интерпретации значения терминов «осмысление» или «осознание». Некоторые отказываются

<sup>27</sup> В.Э.: С этим утверждением, как оно здесь сформулировано у Пенроуза, можно было бы согласиться, если только под «вычислением» понимать работу программы, обрабатывающей информацию (а не что-то действительно вычисляющей). Однако, заглянув немножко вперед, я увидел, что там под этим вариантом *А* Пенроуз понимает уже нечто другое, чем мог бы здесь подразумевать я при этих словах. Поэтому я вынужден отгородиться от всех четырех вариантов Пенроуза, которых, по его мнению, может придерживаться «разумный индивид», и заявить, что я придерживаюсь пятого варианта: *Е* – Всякое мышление есть самопрограммирование, а сознание есть доступность результатов выполнения одних программ для анализа другими программами. (Это немножко упрощено, но точнее и длиннее здесь говорить невозможно).

<sup>28</sup> Например, Д. Деннет, Д. Хофштадтер, М. Мински, Х. Моравек, Г. Саймон; подробнее о терминах можно прочесть в [340], [243].

признавать само существование такого феномена,<sup>29</sup> как «осмысленное осознание», тогда как другие собственно феномен признают, однако рассматривают его лишь как своего рода «эмергентное свойство» (см. также §4.3 и §4.4), которое проявляется всякий раз, когда выполняемое вычисление имеет достаточную степень сложности (или громоздкости, или самоотносимости, или чего угодно еще). В §1.12 я приведу свою собственную интерпретацию терминов «осознание» и «осмысление». Пока же любые расхождения в возможной их интерпретации не будут иметь особой важности для наших рассуждений.

Аргументы, приведенные мной в НРК, были направлены, главным образом, против точки зрения  $\mathcal{A}$ , или позиции сильного ИИ. Один только объем этой книги должен показать, что, хотя лично я не верю в истинность  $\mathcal{A}$ , я всё же рассматриваю эту точку зрения как реальную возможность, на которую стоит обратить серьезное внимание.  $\mathcal{A}$  есть следствие предельно операционного подхода к науке, предполагающего, что абсолютно все феномены физического мира можно описать одними лишь вычислительными методами. В одной из крайних вариаций такого подхода сама вселенная рассматривается, по существу, как единый гигантский компьютер,<sup>30</sup> причем «осмысленные осознания», формирующие, в сущности, наш с вами сознательный разум, вызываются посредством соответствующих субвычислений, выполняемых этим компьютером.

Я полагаю, что эта точка зрения (согласно которой физические системы следует считать простыми вычислительными объектами) отчасти основывается на значительной и постоянно растущей роли вычислительных моделей в современной науке и отчасти на убеждении в том, что сами физические объекты – это, в некотором смысле, всего лишь «информационные модели», подчиняющиеся математическим, вычислительным законам. Большая часть материи, из которой состоят наше тело и мозг, постоянно обновляется – неизменными остаются лишь их модели. Более того, и сама материя, судя по всему, ведет преходящее существование, поскольку ее можно преобразовать из одной формы в другую. Даже масса материального тела, которая является точной физической мерой количества материи, содержащегося в теле, может быть при определенных обстоятельствах превращена в чистую энергию (в соответствии со знаменитой формулой Эйнштейна  $E = mc^2$ ). Следовательно, и материальная субстанция, по-видимому, способна превращаться в нечто, обладающее лишь теоретико-математической реальностью. Более того, если верить квантовой теории, материальные частицы – это не что иное, как информационные «волны». (На этих вопросах мы более подробно остановимся во второй части книги.) Таким образом, сама материя есть нечто неопределенное и недолговечное, поэтому вполне разумно предположить, что постоянство человеческого «я», возможно, больше связано с сохранением моделей, нежели реальных частиц материи.<sup>31</sup>

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

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

<sup>29</sup> В.Э.: Под словами «осмысленное осознание» (и подобными) в бытовой речи (а «научная речь», к сожалению, здесь не отличается от бытовой) в разных контекстах подразумевают разные вещи. Поэтому бессмысленно «отрицать существование феномена»; надо разбираться, что под этими словами скрывалось в каждом конкретном случае (контексте). Но как единая какая-то «сущность», «осознание», конечно же, не существует.

<sup>30</sup> См. [267].

<sup>31</sup> В.Э.: Ну разумеется – человек (его «душа») – это «набор файлов», и ничего более.

алгоритмами, то мы имеем полное право допустить, что в рамках рассматриваемой ситуации такая модель должна содержать и все внутренние проявления разума (включая собственно сознание).<sup>32</sup>

Принимая или отвергая такой вывод из вышеприведенного рассуждения, которое в основе своей составляет суть так называемого теста Тьюринга,<sup>33</sup> мы тем самым определяем свою принадлежность к тому или иному лагерю – именно здесь проходит граница между позициями *A* и *B*. Согласно *A*, любого управляемого компьютером робота, который после достаточно большого количества заданных ему вопросов ведет себя так, словно он обладает сознанием, следует фактически считать обладающим сознанием.<sup>34</sup> Согласно *B*, робот вполне может вести себя точно так же, как обладающий сознанием человек, при этом реально не имея и малой доли этого внутреннего качества. И *A*, и *B* сходятся в том, что управляемый компьютером робот может вести себя так, как ведет себя обладающий сознанием человек, *C* же, напротив, не допускает и малейшей возможности того, что когда-либо может быть реализована эффективная модель обладающего сознанием человека в виде управляемого компьютером робота. Таким образом, согласно *C*, после некоторого достаточно большого количества вопросов реальное отсутствие сознания у робота так или иначе проявится.<sup>35</sup> Вообще говоря, *C* является в гораздо большей степени операционной точкой зрения, нежели *B*, и в этом отношении она больше похожа на *A*, чем на *B*.

Так что же представляет собой позиция *B*? Я думаю, что *B* – это, вероятно, именно та точка зрения, которую многие полагают «научным здравым смыслом». Описываемый ею искусственный интеллект еще называют слабым (или мягким) ИИ. Подобно *A*, она утверждает, что все физические объекты этого мира должны вести себя в соответствии с некоторыми научными принципами, которые, в принципе, допускают создание вычислительной модели этих объектов. С другой стороны, эта точка зрения уверенно отрицает мнение операционистов, согласно которому любой объект, внешне проявляющий себя как сознательное существо, непременно обладает сознанием. Как отмечает философ Джон Серл,<sup>36</sup> вычислительную модель физического процесса никоим образом не следует отождествлять с самим процессом, происходящим в действительности. (Компьютерная модель, например, урагана – это совсем не то же самое, что и реальный ураган!) Согласно взгляду *B*, наличие или отсутствие сознания очень

---

<sup>32</sup> В.Э.: Это проблема искусственная. Если компьютер не будет иметь «внутренних проявлений разума» (т.е. тех информационных структур, которые это обеспечивают), то его очень скоро «разоблачат» в таком диалоге. Но нет проблем (принципиальных) эти структуры в него встроить.

<sup>33</sup> [369]; см. также НРК, с. 5–14.

<sup>34</sup> В.Э.: Вот это и есть то место, которое выше не позволило мне причислить себя к *A* и заставило объявить себя принадлежащим к пятой точке зрения *E*. Если Пенроуз ТАК определяет *A*, то я не принадлежу к этой группе! Как бы удачно компьютер ни имитировал разумную деятельность, он не будет обладать разумом до тех пор, пока в него не встроит те структуры, которыми разум определяется. В сущности здесь самая основная развилка! ВСЁ! Мы с Пенроузом разошлись – и никогда уже не сойдемся! Теперь Пенроуз испишет горы бумаги, доказывая, что точка зрения *A* ошибочна (ну конечно, ошибочна! – какой дурак может ее держаться?! – во всяком случае уж не я), и Пенроуз даже и краем пальчика не коснется точки зрения *E* (которая и есть Веданская теория), потому что Пенроуз даже не подозревает о ее существовании; он никогда ее не видел, никогда о ней не слышал, никогда ее не разбирал и уж, конечно, не понимает ее... (Скучно стало, господа...)

<sup>35</sup> В.Э.: Разумеется, проявится! И это точка зрения *E*. Но *C* «не допускает и малейшей возможности того, что когда-либо может быть реализована эффективная модель обладающего сознанием человека в виде управляемого компьютером робота», а *E* просто знает, КАК это сделать. Разбирая Первую книгу Пенроуза (НРК) для своих латышских оппонентов (профессора Тамберга и др.), я ввел (весьма важные) понятия имитации и реализации интеллекта. Реализация интеллекта – это когда в компьютер (или любой другой субъект) действительно встраивают все те структуры, которые необходимы для существования разума. Имитация интеллекта – это когда БЕЗ построения этих структур пытаются подражать разумной деятельности. Пенроузовская группа *A* – это люди, считающие, что достаточно хорошая имитация уже и есть интеллект. Пенроузовская группа *C* – это люди, считающие, что реализации интеллекта на компьютерах вообще невозможны. Но, доказывая свой тезис, Пенроуз рассматривает одни лишь имитации – и в конце концов доказывает только одно: что имитация – это не реализация (кто будет спорить против этого?). А настоящие реализации разума Пенроуз просто-напросто вообще не рассматривает – в первую очередь потому, что он просто не знает, как реализация должна выглядеть, и чем она отличается от имитации. **В этом вся суть:** Пенроуз разбирается ТОЛЬКО с имитациями.

<sup>36</sup> См. [340], [341].

сильно зависит от того, какой именно физический объект «осуществляет мышление» и какие физические действия он при этом совершает. И только потом следует рассмотреть конкретные вычисления, которых требуют эти действия. Таким образом, активность биологического мозга может вызвать осознание, а вот его точная электронная модель вполне может оказаться на это неспособной. Это различие, по  $\mathcal{B}$ , совсем не обязательно должно оказаться различием между биологией и физикой. Однако крайне важным остается реальное материальное строение рассматриваемого объекта (скажем, мозга), а не просто его вычислительная активность.

Позиция  $\mathcal{C}$ , на мой взгляд, ближе всех к истине. Она подразумевает более операционный подход, нежели  $\mathcal{B}$ , так как утверждает, что существуют такие внешние проявления обладающих сознанием объектов (скажем, мозга), которые отличаются от внешних проявлений компьютера: внешние проявления сознания невозможно должным образом воспроизвести вычислительными методами. Свои основания для такой убежденности я приведу несколько позже. Поскольку  $\mathcal{C}$ , как и  $\mathcal{B}$ , не отвергает позиции физикалистов, согласно которой разум возникает в результате проявления активности тех или иных физических объектов (например, мозга, хотя это и не обязательно), следовательно, подразумевает, что не всякую физическую активность можно должным образом смоделировать вычислительными методами.<sup>37</sup>

Допускает ли современная физика возможность существования процессов, которые принципиально невозможно смоделировать на компьютере? Если мы надеемся получить на этот вопрос математически строгий ответ, то нас ждет разочарование. По крайней мере, лично мне такой ответ неизвестен. Вообще, с математической точностью здесь дело обстоит несколько запутаннее, чем хотелось бы.<sup>38</sup> Однако сам я убежден в том, что подобные невычислимые процессы следует искать за пределами тех областей физики, которые описываются известными на настоящий момент физическими законами. Далее в этой книге я вновь перечислю некоторые весьма серьезные – причем именно физические – доводы в пользу того, что мы действительно нуждаемся в новом взгляде на ту область, которая лежит между уровнем микроскопических величин, где господствуют квантовые законы, и уровнем «обычных» размеров, подвластным классической физике. Хотя, надо сказать, далеко не все современные физики единодушно уверены в необходимости подобной новой физической теории.

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

Кое-кто попытался было объявить, что эти соображения отправляют меня прямоком в лагерь сторонников точки зрения  $\mathcal{D}$ , поскольку я утверждаю, что для отыскания хоть какого-то объяснения феномену сознания нам придется выйти за пределы известной науки. Однако между упомянутым строгим вариантом  $\mathcal{C}$  и точкой зрения  $\mathcal{D}$  есть существенная разница, в частности, на уровне методологии. В соответствии с  $\mathcal{C}$ , проблема осмысленного осознания носит, в сущности, научный характер, даже если подходящей наукой мы пока что не располагаем. Я всецело поддерживаю эту точку зрения; я полагаю, что ответы на интересующие нас вопросы нам следует искать именно с помощью научных методов – разумеется, должным образом усовершенствованных, пусть даже о конкретной природе необходимых изменений мы, возможно, имеем на

<sup>37</sup> В.Э.: Здесь вообще раскрываются некоторые (ошибочные с нашей точки зрения) представления Пенроуза. Для него компьютер моделирует мозг. (И тогда не все аспекты работы мозга можно смоделировать, и т.д. и т.п...) На самом деле компьютер (в который встроен разум) ничего не моделирует. Компьютер и мозг просто делают одну и ту же работу: обрабатывают информацию о внешней среде и о своем собственном состоянии. И обрабатывают по одинаковым законам информатики.

<sup>38</sup> Вопрос осложняется тем, что современная физика рассматривает, по большей части, непрерывные, а не дискретные (цифровые) процессы. Самый смысл термина «вычислимость» в данном контексте можно трактовать по-разному. С некоторыми рассуждениями на данную тему можно ознакомиться в [312], [346], [313], [314], [315], [316], [30], [327], [328]. К этому вопросу я еще вернусь в §1.8.

данный момент лишь самое смутное представление. В этом и состоит ключевая разница между  $\mathcal{C}$  и  $\mathcal{D}$ , насколько бы похожими не казались нам соответствующие мнения относительно того, на что способна современная наука.

Определенные выше точки зрения  $\mathcal{A}$ ,  $\mathcal{B}$ ,  $\mathcal{C}$ ,  $\mathcal{D}$  представляют собою крайности, или полярные точки возможных позиций, которых может придерживаться тот или иной индивидуум. Я вполне допускаю, что кому-то может показаться, что их собственные взгляды не подходят ни под одну из перечисленных категорий, а лежат где-то между ними либо противоречат некоторым из них. Безусловно, между такими, например, крайними точками зрения, как  $\mathcal{A}$  и  $\mathcal{B}$ , можно разместить множество различных промежуточных точек зрения (см. [344]). Существует даже мнение (весьма, кстати, широко распространенное), которое лучше всего определяется как комбинация  $\mathcal{A}$  и  $\mathcal{D}$  (или, быть может,  $\mathcal{B}$  и  $\mathcal{D}$ ), – предусматриваемая им возможность еще сыграет немаловажную роль в наших дальнейших размышлениях. Согласно этому мнению, мозг действительно работает как компьютер, однако компьютер настолько невообразимой сложности, что его имитация не под силу человеческому и научному разумению, ибо он, несомненно, является божественным творением Господа – «лучшего в мире системотехника»<sup>39</sup>, не иначе!

#### §1.4. Физикализм и ментализм

Я должен сделать здесь краткое отступление касательно использования терминов «физикалист» и «менталист», обычно противопоставляемых один другому, в нашей конкретной ситуации, т.е. в отношении крайних точек зрения, обозначенных нами через  $\mathcal{A}$ ,  $\mathcal{B}$ ,  $\mathcal{C}$  и  $\mathcal{D}$ . Поскольку  $\mathcal{D}$  являет собой полное отрицание физикализма, сторонников  $\mathcal{D}$  безусловно следует считать менталистами. Однако мне не совсем ясно, где провести границу между физикализмом и ментализмом в случае с тремя другими позициями  $\mathcal{A}$ ,  $\mathcal{B}$  и  $\mathcal{C}$ . Я полагаю, что приверженцев  $\mathcal{A}$  следует обыкновенно считать физикалистами, и я уверен, что подавляющее их большинство согласилось бы со мной. Однако здесь скрывается некий парадокс. В соответствии с  $\mathcal{A}$ , материальное строение мыслящего устройства считается несущественным. Все его мыслительные атрибуты определяются лишь вычислениями, которые это устройство выполняет. Сами по себе вычисления суть феномены абстрактной математики,<sup>40</sup> не связанные с конкретными материальными телами. Таким образом, согласно  $\mathcal{A}$ , сами мыслительные атрибуты не имеют жесткой связи с физическими объектами, а потому термин «физикалист» может показаться несколько неуместным. Точки зрения  $\mathcal{B}$  и  $\mathcal{C}$ , напротив, требуют, чтобы при определении наличия в том или ином объекте подлинного разума решающую роль играло реальное физическое строение рассматриваемого объекта. Соответственно, вполне можно было бы утверждать, что именно эти точки зрения, а никак не  $\mathcal{A}$ , представляют возможные позиции физикалистов. Однако такая терминология, по-видимому, вошла бы в некоторое противоречие с общепринятым употреблением, где более уместным считается называть «менталистами» сторонников  $\mathcal{B}$  и  $\mathcal{C}$ , поскольку в этих случаях свойства мышления рассматриваются как нечто «реальное», а не просто как «эпифеномены»<sup>41</sup>, которые случайным образом возникают при выполнении определенных типов вычислений. Ввиду такой путаницы, я буду избегать использования терминов «физикалист» и «менталист» в последующих рассуждениях, ссылаясь вместо этого на конкретные точки зрения  $\mathcal{A}$ ,  $\mathcal{B}$ ,  $\mathcal{C}$  и  $\mathcal{D}$ , определенные выше.

#### §1.5. Вычисление: нисходящие и восходящие процедуры

До сих пор было не совсем ясно, что именно я понимаю под термином «вычисление» в определениях позиций  $\mathcal{A}$ ,  $\mathcal{B}$ ,  $\mathcal{C}$  и  $\mathcal{D}$  приведенных в §1.3. Что же такое вычисление? В двух словах: это всё, что делает самый обычный универсальный компьютер. Если же мы хотим быть более

<sup>39</sup> Этой замечательной фразой я обязан диктору BBC Radio 4, ведущему программу «Мысль дня».

<sup>40</sup> В.Э.: Ух-х – бац! Нифига себе заявление! (На самом деле компьютерная деятельность – обработка информации – никакого отношения к абстрактной математике не имеет. Наоборот, вся «абстрактная математика» является отдаленной «тенью» компьютерной деятельности. Компьютеры – первичны; математика вторична – или даже третьична).

<sup>41</sup> Эпифеномен – побочное явление, сопутствующее другим явлениям (феноменам), но не оказывающее на них никакого влияния. – *Прим. перев.*

точными, то следует воспринимать этот термин в соответственно идеализированном смысле: вычисление – это действие машины Тьюринга.

А что такое машина Тьюринга? По сути, это и есть математически идеализированный компьютер (теоретический предшественник современного универсального компьютера); идеализирован же он в том смысле, что никогда не ошибается, может работать сколько угодно долго и обладает неограниченным объемом памяти. Немного более подробно о точных спецификациях машин Тьюринга я расскажу в §2.1 и в Приложении А (с. 191). (Интересующийся более полным введением в этот вопрос читатель может обратиться к описанию, приведенному в НРК, глава 2, а также к работам Клина [223] или Дэвиса [72].)

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

Мы говорим, что вычислительная процедура имеет нисходящую организацию, если она построена в соответствии с некоторой прозрачной и хорошо структурированной фиксированной вычислительной процедурой (которая может содержать некий заданный заранее объем данных) и предоставляет, в частности, четкое решение для той или иной рассматриваемой проблемы. (Описанный в НРК на с.31<sup>42</sup> евклидов алгоритм нахождения наибольшего общего делителя двух натуральных чисел представляет собой простой пример нисходящего алгоритма.) В противоположность такой организации существует организация восходящая, где упомянутые четкие правила выполнения действий и объем данных заранее не определены, однако вместо этого имеется некоторая процедура, определяющая, каким образом система должна «обучаться» и повышать свою эффективность в соответствии с накопленным «опытом». Иными словами, в случае восходящей системы правила выполнения действий подвержены постоянному изменению. Очевидно, что такая система должна пройти множество циклов, выполняя требуемые действия над непрерывно поступающими данными. Во время каждого прогона производится оценка эффективности (возможно, самой системой), после чего, в соответствии с этой оценкой, система так или иначе модифицирует свои действия, стремясь улучшить качество вывода данных. Например, на вход системы подаются несколько оцифрованных с некоторым качеством фотопортретов, и ставится задача – определить, на каких портретах изображен один человек, а на каких – другой. После каждого прогона результат выполнения задачи сравнивается с правильным, после чего правила выполнения действий модифицируются так, чтобы с некоторой вероятностью добиться улучшения функционирования системы при следующем прогоне.<sup>43</sup>

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

Для наших целей важно понимать, что и нисходящие, и восходящие вычислительные процедуры с легкостью выполняются на универсальном компьютере, а потому их можно отнести к категории процессов, названных мною вычислительными и алгоритмическими. Таким образом, в случае восходящих (или комбинированных) систем сам способ модификации системой своих

---

<sup>42</sup> Напомним, что здесь и далее приводятся страницы оригинального английского издания. – *Прим. перев.*

<sup>43</sup> В.Э.: Ну-у... Это НЕ самопрограммирование! При ТАКИХ программах интеллект не построят! И не велика заслуга это доказать.

процедур задается какими-то целиком и полностью вычислительными инструкциями, причем задается заблаговременно. Этим и объясняется возможность реализации всей системы на обычном компьютере. Существенная разница между восходящей (или комбинированной) системой и системой нисходящей состоит в том, что в первом случае вычислительная процедура должна подразумевать возможность сохранения «памяти» о предыдущем выполнении задачи (т.е. обладать способностью накапливать «опыт») с тем, чтобы эту память затем можно было использовать в последующих вычислительных действиях. Конкретные подробности сейчас не имеют особого значения, однако к обсуждению этого вопроса мы еще вернемся в [§3.11](#).

Задавшись целью создать искусственный интеллект (сокращенно «ИИ»), человек пока лишь пытается симитировать разумное поведение на каком угодно уровне посредством каких-то вычислительных средств. При этом часто используется как нисходящая, так и восходящая организация. Первоначально наиболее перспективными представлялись нисходящие системы,<sup>44</sup> однако сейчас всё большую популярность приобретают восходящие системы типа искусственной нейронной сети. По всей видимости, получения наиболее успешных систем ИИ можно ожидать лишь при том или ином сочетании нисходящих и восходящих организаций. У каждой из них есть свои преимущества. Нисходящая организация наиболее успешна в тех областях, где данные и правила выполнения действий четко определены и имеют хорошо выраженный вычислительный характер – при решении некоторых конкретных математических задач, создании вычислительных систем для игры в шахматы или, скажем, в медицинской диагностике, где определение того или иного заболевания происходит с помощью заданных наборов правил, основанных на общепринятых медицинских процедурах. Восходящая же организация оказывается полезной, когда критерии для принятия решений не слишком точны или не совсем ясны – как, например, при распознавании лиц или звуков или, возможно, при поиске месторождений минералов, где основным поведенческим критерием становится повышение эффективности на основе накопленного опыта. Во многих подобных системах действительно присутствуют элементы и нисходящей, и восходящей организаций (например, шахматный компьютер, обучающийся на основе опыта, или созданное на базе какой-либо четкой геологической теории вычислительное устройство, помогающее в поисках месторождений минералов).

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

Еще одно отличие между видами компьютерных систем связано с различием между последовательной и параллельной архитектурами. Компьютер последовательного действия – это машина, выполняющая вычисления друг за другом, поэтапно, тогда как параллельный компьютер выполняет множество независимых вычислений одновременно, результаты же этих вычислений сводятся вместе лишь по завершении достаточно большого их количества. Причем у истоков разработки некоторых параллельных систем стояли все те же теории, описывающие

---

<sup>44</sup> Исследования в области создания ИИ начались в 1950-е годы с весьма успешного применения сравнительно элементарных нисходящих процедур (например, Грей Уолтер, 1953). Распознающий образы «перцептрон» Фрэнка Розенблатта [323] стал в 1959 году первым удачным «связным» устройством (искусственной нейронной сетью), вызвав тем самым значительный интерес к схемам восходящего типа. В 1969 году Марвин Мински и Сеймур Пейперт указали на некоторые существенные ограничения, присущие данному типу восходящей организации (см. [264]). Способ обойти эти ограничения предложил некоторое время спустя Хопфилд [207], и в настоящий момент искусственными устройствами, функционирующими по типу нейронной сети, активно занимаются ученые всего мира. (О применении таких устройств, например, в физике высоких энергий см. [20] и [142].) Что касается ИИ нисходящего типа, то здесь важными вехами стали работы Джона Маккарти [248] и Алана Ньюэлла в сотрудничестве с Гербертом Саймоном [272]. Впечатляющее изложение истории исследований проблемы ИИ можно найти в [124]. Из прочей литературы порекомендую [175], [16] (относительно недавние размышления о процедурах и перспективах ИИ); [98] (классическая критика идеи ИИ); [140] (свежий взгляд на проблему от пионера ИИ); также см. статьи в сборниках [41] и [221].

предполагаемые способы функционирования мозга. Здесь следует отметить, что различие между вычислительными машинами последовательного и параллельного действия ни в коей мере не является принципиальным. Параллельное действие всегда можно смоделировать последовательно,<sup>45</sup> хотя, конечно же, существуют некоторые типы задач (весьма немногочисленные), для решения которых эффективнее (в смысле затрат времени на вычисление и т.п.) будет параллельное действие, нежели последовательное. Поскольку в рамках настоящего труда меня занимают, главным образом, принципиальные вопросы, различия между параллельными и последовательными вычислениями не представляются в этом отношении особенно существенными.

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

Вспомним, что точка зрения  $\mathcal{C}$  предполагает, что обладающий сознанием мозг функционирует таким образом, что его активность не поддается никакому численному моделированию – ни нисходящего, ни восходящего, ни какого-либо другого типа. Те, кто сомневается в истинности  $\mathcal{C}$  могут отчасти оправдать свои сомнения тем, что формулировка  $\mathcal{C}$  якобы противоречит так называемому тезису Черча (или тезису Черча–Тьюринга) – вернее, тому условию, которое сейчас общепринято обозначать упомянутым термином. В чем же суть тезиса Черча? В первоначальной форме, предложенной американским логиком Алонзо Черчем в 1936 году, этот тезис гласил, что любой процесс, который можно корректно назвать «чисто механическим» математическим процессом, – т.е. любой алгоритмический процесс – может быть реализован в рамках конкретной схемы, открытой самим Черчем и названной им лямбда-исчислением ( $\lambda$ -исчислением)<sup>46</sup> (весьма, надо отметить, изящная и концептуально сдержанная схема; краткое ознакомительное изложение см. в НРК, с. 66–70). Вскоре после этого, в 1936–1937 годах, британский математик Алан Тьюринг нашел свой собственный, гораздо более убедительный способ описания алгоритмических процессов, основанный на функционировании теоретических «вычислительных машин», которые мы сейчас называем машинами Тьюринга. Вслед за Тьюрингом в некоторой степени аналогичную схему разработал американский ученый-логик польского происхождения Эмиль Пост (1936). Далее Черч и Тьюринг независимо друг от друга показали, что исчисление Черча эквивалентно концепции машины Тьюринга (а следовательно, и схемы Поста). Более того, именно этим концепциям Тьюринга в значительной степени обязаны своим появлением на свет современные универсальные компьютеры. Как уже упоминалось, машина Тьюринга по принципу функционирования фактически полностью эквивалентна современному компьютеру, – несколько, впрочем, идеализированному, т.е. обладающему возможностью использовать неограниченный объем памяти. Таким образом получается, что тезис Черча в его первоначальной формулировке всего лишь утверждает, что математическими алгоритмами следует считать как раз те процессы, которые способен выполнить идеализированный современный компьютер – а если учесть общепринятое ныне определение термина «алгоритм», то такое утверждение и вовсе становится тавтологией. Так что принятие этой формулировки тезиса Черча не влечет за собой никакого противоречия точке зрения  $\mathcal{C}$ .<sup>47</sup>

Вполне вероятно, однако, что сам Тьюринг имел в виду нечто большее: вычислительные возможности любого физического устройства должны (в идеале) быть эквивалентны действию машины Тьюринга. Такое утверждение существенно выходит за рамки того, что изначально подразумевал Черч. При разработке концепции «машины Тьюринга» сам Тьюринг основывался

---

<sup>45</sup> В.Э.: Это утверждение Пенроуза ошибочно. (Видимо, наслушался «великих теоретиков» по «машинам Тьюринга»). На линейном (последовательном) компьютере никакой реальный искусственный интеллект не сделаешь – что бы там ни говорили «теоретики». Параллельность (и независимость) различных процессов является необходимой частью программы ИИ. (Возможно, эта уверенность в эквивалентности параллельных и последовательных процессов помогает Пенроузу не понимать, как можно сделать реализацию – а не имитацию – разума).

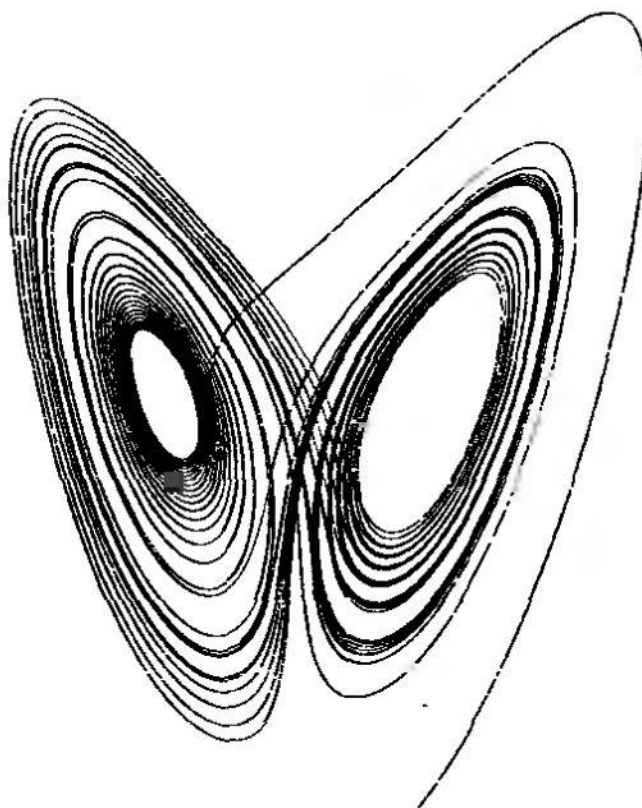
<sup>46</sup> Описание лямбда-исчисления см. в [53] и [223].

<sup>47</sup> Время от времени математики натываются на процедуру, которая «очевидно» алгоритмична по своей природе, пусть даже порой не всегда бывает ясно, как эту процедуру можно сформулировать в виде операций машины Тьюринга или лямбда-исчисления. В таких случаях можно утверждать, что, «согласно тезису Черча», такая операция и в самом деле должна существовать. См., например, [67]. В этом пути нет ничего зазорного, и, уж конечно, не возникает никакого противоречия с  $\mathcal{C}$ . Более того, на таком толковании тезиса Черча основывается большая часть рассуждений главы 3.

на своих представлениях о том, чего, в принципе, мог бы достичь вычислитель-человек (см. [198]). Судя по всему, он полагал, что физическое действие в общем (а под эту категорию подпадает и активность мозга человека) всегда можно свести к какой-либо разновидности действия машины Тьюринга. Быть может, это утверждение (физическое) следует называть «тезисом Тьюринга» – для того чтобы отличать его от оригинального «тезиса Черча», утверждения чисто математического, которому никоим образом не противоречит  $\mathcal{C}$ . Именно такой терминологии я намерен придерживаться далее в этой книге. Соответственно, точка зрения  $\mathcal{C}$  противоречит в этом случае тезису Тьюринга, а вовсе не тезису Черча.

### §1.7. Хаос

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



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

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

которая может оказаться просто недостижимой не только для тех измерительных средств, которыми мы располагаем, но также и для тех, которые мы только можем вообразить.

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

С другой стороны, подобное – так называемое хаотическое – поведение может иметь место и в случае очень простых систем; примером тому служат системы, состоящие из малого количества частиц. Вообразите, что от вас требуется загнать в лузу бильярдный шар Е, расположенный пятым в некоторой извилистой<sup>48</sup> и очень растянутой цепочке шаров А, В, С, D и Е; вам нужно ударить кием по шару А так, чтобы тот ударил шар В, который, в свою очередь, ударил бы шар С, который ударил бы шар D, который ударил бы шар Е, который, наконец, попал бы в лузу. В общем случае необходимая для этого точность значительно превышает способности любого профессионального игрока в бильярд. Если бы цепочка состояла из 20 шаров, то тогда – даже допустив, что эти шары представляют собой идеально упругие точные сферы – задача загнать в лузу последний шар оказалась бы не под силу и самому точному механизму из всех доступных современной технологии. Поведение последних шаров цепочки было бы, в сущности, случайным, несмотря на то, что управляющие поведением шаров ньютоновы законы математически абсолютно детерминированы и, в принципе, эффективно вычислимы. Никакое вычисление не смогло бы предсказать реальное поведение последних шаров цепочки просто потому, что нет никакой возможности добиться достаточно точного определения реального начального положения и скорости движения кия или положений первых шаров цепочки. Более того, даже самые незначительные внешние воздействия, вроде дыхания человека в соседнем городе, могут нарушить эту точность до такой степени, которая полностью обесценит результаты любого подобного вычисления.

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

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

---

<sup>48</sup> В черновом варианте книги слова «извилистой» здесь не было. Если шары расположены точно на прямой линии, этот трюк оказывается достаточно простым: я узнал об этом, к своему удивлению, когда попробовал проделать это сам. При расстановке шаров по прямой возникает неожиданная устойчивость, отсутствующая в общем случае.

природу таких операций: если запустить процесс компьютерного моделирования вторично, задав те же входные данные, что и ранее, то результат моделирования будет точно таким же, как и в первый раз! (Здесь предполагается, что сам компьютер не ошибается; впрочем, надо признать, что современные компьютеры и в самом деле крайне редко совершают при вычислениях реальные ошибки.)

Возвращаясь к искусственному интеллекту, отметим, что никто пока и не пытается воспроизвести поведение какого-то конкретного индивидуума<sup>49</sup>; нас бы прекрасно устроила модель индивидуума вообще<sup>50</sup>! В этом контексте моя позиция вовсе не представляется такой уж неразумной: хаотические системы следует безусловно относить к категории систем, которые мы называем «вычислительными». Компьютерная модель такой системы и в самом деле выглядела бы как абсолютно приемлемый «типичный случай», даже и не совпадая при этом ни с каким «реальным случаем». Если внешние проявления человеческого разума суть результаты некоей хаотической динамической эволюции (эволюции вычислительной в том смысле, о котором мы только что говорили), то это вполне согласуется с точками зрения *A* и *B*, но *никак не C*.

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

### §1.8. Аналоговые вычисления

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

В случае аналоговых систем необходимо учитывать одно формальное обстоятельство: стандартные понятия вычисления и вычислимости применимы, строго говоря, только к дискретным системам (над которыми, собственно, и выполняются «цифровые» действия), но не к

---

<sup>49</sup> В.Э.: Вообще здесь опять проявляется непонимание Пенроузом самых основных вещей ИИ. У него в голове стоит только имитация, а не реализация интеллекта. «*Не пытается воспроизвести поведение какого-то конкретного индивидуума...*» В том-то и дело, что если на компьютере будет создан полноценный интеллект, то это будет уникальная личность – так же, как уникален каждый человек на Земле. И если на другом компьютере будет создан второй искусственный интеллект, то это будет опять уникальная личность, не совпадающая с первой (если только специально не позаботятся о том, чтобы они были идентичны, запуская в них 100% идентичные стартовые программы и поставляя потом им 100% идентичную информацию; в природе 100% идентичные стартовые программы имеют однойцевые близнецы, но дальнейшая информация у них всё равно отличается и никогда не совпадает полностью).

<sup>50</sup> В.Э.: «Модель индивидуума вообще» может быть только теоретической – аналогом учебника «Анатомия и физиология», только выполненным не для тела, а для «духа». Веданская теория как раз и есть такая «теоретическая модель индивидуума вообще». А конкретные реализации «разума» могут быть только уникальными личностями, как и люди. (Видно, что всё время Пенроуз думает только об имитациях, только об имитациях интеллекта, а не о его реализациях!)

непрерывным,<sup>51</sup> таким, например, как расстояния или электрические потенциалы, с которыми имеет дело традиционная классическая физика. Иными словами, для того чтобы применить обычные вычислительные понятия к системе, описание которой требует не дискретных (или «цифровых»), а непрерывных параметров, мы естественным образом должны прибегнуть к аппроксимации. Действительно, при компьютерном моделировании физических систем<sup>52</sup> вообще стандартной процедурой является аппроксимация всех рассматриваемых непрерывных параметров в дискретной форме. Подобная процедура, однако, неминуемо вносит некоторую погрешность, величина которой определяется заданной степенью точности аппроксимации; при этом вполне возможно, что для той или иной интересующей нас физической системы заданной точности может оказаться недостаточно. В итоге дискретное компьютерное моделирование очень просто может привести нас к ошибочным выводам относительно поведения моделируемой непрерывной физической системы.

В принципе, ничто не мешает повысить точность до уровня, адекватного для моделирования рассматриваемой непрерывной системы. Однако на практике, особенно в случае хаотических систем, требуемые для этого время вычислений и объем памяти могут оказаться непомерно большими. Кроме того, можем ли мы, строго говоря, быть абсолютно уверенными в том, что выбранная нами степень точности является действительно достаточной. Необходим какой-то критерий, который позволил бы нам определить, что нужный уровень точности достигнут, дальнейшего ее повышения не требуется и качественному поведению, вычисленному с такой точностью, в самом деле можно доверять. Всё это поднимает ряд достаточно щекотливых математических вопросов, рассматривать которые подробно на этих страницах мне представляется не совсем уместным.

Существуют, однако, и другие подходы к проблемам вычислений в случае непрерывных систем; например, такие, в которых непрерывные системы рассматриваются как самостоятельные математические структуры со своим собственным понятием «вычислимости» – понятием, обобщающим идею вычислимости по Тьюрингу с дискретных величин на непрерывные.<sup>53</sup> При таком подходе исчезает необходимость в аппроксимации непрерывной системы дискретными параметрами с целью применить к ней традиционную концепцию вычислимости по Тьюрингу. Такие идеи вызывают определенный интерес с математической точки зрения; к сожалению, им, как нам представляется, не достаёт пока той неотразимой естественности и уникальности, которые присущи стандартному понятию вычислимости по Тьюрингу для дискретных систем. Более того, вследствие определенной непоследовательности данного подхода, формально «невыхислимыми» оказываются и некоторые простые системы, в применении к которым подобная терминология выглядит как-то не совсем уместно (даже такие, например, как известное всем из физики простое «волновое уравнение»; см. [314] и НРК, с. 187–188). С другой стороны, следует упомянуть и об одной сравнительно недавней работе ([328]), в которой показано, что теоретические аналоговые компьютеры, объединяемые в некоторый достаточно обширный класс, не могут выйти за рамки обычной вычислимости по Тьюрингу. Я надеюсь, что дальнейшие исследования должным образом осветят эти безусловно интересные и важные темы. Пока же у меня нет оснований полагать, что работы в этом направлении в целом уже достигли той стадии завершенности, чтобы их результаты можно было применить к рассматриваемым здесь проблемам.

---

<sup>51</sup> В.Э.: Разница между «дискретными» и «аналоговыми» системами не так уж и велика. Рассматривая хотя бы ту же логарифмическую линейку под всё большим и большим увеличением, мы, начиная с какого-то момента, обнаружим, что ее «непрерывные» грани исчезли, и вместо них имеются в наличии «дискретные» молекулы, а разбирая всё глубже электрическое напряжение, мы дойдем в конце концов до квантовых скачков... С другой стороны, если взять гигантский массив (допустим, миллиарды терабайтов) обычной бинарной информации и вывести ее на подходящий гигантский экран, то видимые там линии считать непрерывными у нас будет не меньше оснований, чем те причины, по которым мы считаем непрерывными края логарифмической линейки.

<sup>52</sup> В.Э.: Только вокруг моделирования, только вокруг имитации крутятся мысли Пенроуза... Но задача у нас не моделирование мозга, а создание системы, которая будет делать ту же работу. Изучать надо эту работу, и думать надо, как ее можно сделать, а не пускаться в рассуждения о моделировании!

<sup>53</sup> Из различных публикаций, посвященных данной проблематике, могу порекомендовать, например, [312], [346], [316], [30]. Вопрос о функционировании мозга в связи с упомянутыми проблемами рассмотрен, в частности, в [326].

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

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

Даже если не принимать во внимание всевозможные теоретические проблемы общего плана, на сегодняшний день наибольшее превосходство перед аналоговыми вычислительными системами демонстрируют именно цифровые компьютеры.<sup>54</sup> Цифровые вычисления имеют гораздо более высокую точность благодаря, в основном, тому, что при хранении данных в цифровом виде повышение точности обеспечивается простым увеличением разрядности чисел, что легко достижимо с помощью весьма скромного увеличения (логарифмического) мощности компьютера; в аналоговых же машинах (по крайней мере, в полностью аналоговых, в конструкцию которых не заложено никаких цифровых концепций) увеличения точности можно добиться лишь посредством весьма и весьма значительного увеличения (линейного) соответствующих параметров. Возможно, когда-нибудь в будущем возникнут новые идеи, которые пойдут на пользу аналоговым вычислителям, однако в рамках современной технологии большая часть существенных практических преимуществ принадлежит, по всей видимости, цифровому вычислению.

### §1.9. Невычислительные процессы

Из всех типов вполне определенных процессов, что приходят в голову, большая часть относится, соответственно, к категории феноменов, называемых мною «вычислительными» (имеются в виду, конечно же, «цифровые вычисления»). Возможно, читатель уже начал волноваться, что сторонники позиции  $\mathcal{C}$  так и останутся у нас не при деле. Причем я еще ни словом не упоминал о строго случайных процессах, которые могут быть обусловлены, скажем, какими-либо исходными данными, получаемыми от квантовой системы. (О квантовой механике мы немного подробнее поговорим во второй части, главы 5 и 6.) Впрочем, для самой системы практически безразлично, подается на ее вход подлинно случайная последовательность данных

---

<sup>54</sup> В.Э.: Я уже показал, что в смысле прерывности-непрерывности разницы между дигитальными и аналоговыми компьютерами почти нет. Но при этом двоичные компьютеры универсальны (могут решать какие угодно задачи), а аналоговые – узко специализированы (созданы каждый для решения одной-единственной задачи или, в лучшем случае, для нескольких близких задач). Поэтому аналоговые компьютеры и были вытеснены двоичными. (Но когда я в 1960-х годах был студентом, которого обучали компьютерам, нам оба эти класса «вычислительных машин» преподносили как одинаково перспективные).

или же всего лишь псевдослучайная, которую можно целиком и полностью сгенерировать вычислительным путем (см. §3.11). Действительно, несмотря на то, что между «случайным» и «псевдослучайным», строго говоря, существуют некоторые формальные отличия, они, на первый взгляд, не имеют непосредственного отношения к проблемам ИИ. Далее, в §3.11, §3.18 и последующих, я приведу некоторые серьезные доводы в пользу того, что «чистая случайность» и в самом деле абсолютно бесполезна для наших целей; если уж возникает такая необходимость, то лучше всё же придерживаться псевдослучайности хаотического поведения, а все нормальные типы хаотического поведения, как уже подчеркивалось выше, относятся к категории «вычислительных».

А что нам известно о роли окружения? По мере развития каждого индивидуума у него или у нее формируется уникальное окружение, отличное от окружения любого другого человека. Возможно, именно это уникальное личное окружение и дает каждому из нас ту особенную последовательность входных данных, которая неподвластна вычислению? Хотя лично мне, например, сложно сообразить, на что именно в данном контексте может повлиять «уникальность» нашего окружения. Эти рассуждения напоминают разговор о хаосе, который мы вели выше (см. §1.7). Для обучения управляемого компьютером робота достаточно одной лишь модели некоего правдоподобного окружения (хаотического), при том, разумеется, условии, что в этой модели не будет ничего заведомо невычислимого. Роботу нет нужды учиться тем или иным навыкам в каком-то конкретном реальном окружении; его, разумеется, вполне устроит типичное окружение, моделирующее реальность вычислительными методами.<sup>55</sup>

А может быть, численное моделирование пусть даже всего лишь правдоподобного окружения невозможно в принципе. Быть может, в окружающем физическом мире есть-таки нечто такое, что на самом деле неподвластно численному моделированию. Возможно, некоторые сторонники *A* или *B* уже вознамерились приписать все не поддающиеся, на первый взгляд, вычислению проявления человеческого поведения невычислимости внешнего окружения. Должен, однако, заметить, что намерение это несколько опрометчиво. Ибо, как только мы признаем, что физическое поведение допускает где-то что-то такое, что невозможно моделировать вычислительными методами, мы тем самым тут же лишаемся главного, по всей видимости, основания сомневаться в правдоподобии, в первую очередь, самой точки зрения *C*. Если во внешнем окружении (т.е. вне мозга) имеют место процессы, не поддающиеся численному моделированию, то почему не могут оказаться таковыми и процессы, протекающие внутри мозга? В конце концов, внутренняя физическая организация мозга человека, по всей видимости, гораздо более сложна, чем большая часть (и это еще слабо сказано) его окружения, за исключением, быть может, тех его участков, где это окружение само оказывается под сильным влиянием деятельности других мозгов. Признание возможности внешней невычислимой физической активности лишает всякой силы главный аргумент против *C*. (См. также §3.9, §3.10.)

Следует сделать еще одно замечание относительно «не поддающихся вычислению» процессов, возможность существования которых предполагает позиция *C*. Под этим термином я имею в виду отнюдь не те процессы, которые всего-навсего невычислимы практически. Здесь, конечно же, уместно вспомнить и о том, что, хотя моделирование любого правдоподобного окружения, или же любое точное воспроизведение всех физических и химических процессов, протекающих в мозге,<sup>56</sup> может быть, в принципе, вычислимым, на такое вычисление, скорее

---

<sup>55</sup> В.Э.: И насчет «случайности», и насчет «обучения роботов» Пенроуз блуждает по совершенно неверным тропинкам, но разбирательство со всем этим здесь заняло бы так много места, что отложим его на потом.

<sup>56</sup> В.Э.: Ну так!... Здесь уже Пенроуз пошел «не в ту степь» до такой степени, что, пожалуй, и можно дальше не читать (разве что из чистого интереса: что еще он наговорит?!). Он же вообще глобально дезориентирован! Ну какое «моделирование любого правдоподобного окружения», ну какое «точное воспроизведение всех физических и химических процессов, протекающих в мозге»? О чем он здесь говорит?? Всё это не имеет НИКАКОГО отношения к искусственному разуму. При построении искусственного разума (интеллекта) мы имеем следующую задачу: 1) определить, какую работу выполняет человеческий мозг; и 2) выполнить эту работу на другом устройстве. Это точно так же, как при задаче создания «искусственного землекопа»: 1) определяем, какую работу выполняет землекоп (копает яму); 2) создаем машину (экскаватор), которая тоже копает яму. Вот и всё! И нужно быть полным идиотом (или Оксфордским профессором? ☺), чтобы при конструировании экскаватора начать рассуждать про «точное воспроизведение всех физических и химических процессов, протекающих в мышцах» землекопа и про «моделирование любого правдоподобного грунта!» Очевидно же, что Пенроуз не видит, не ставит и не

всего, понадобится столько времени или такой объем памяти, что вряд ли удастся выполнить его на любом реально существующем или даже воображимом в ближайшем будущем компьютере. Вероятно, нереально даже написание соответствующей компьютерной программы, если учесть, какое огромное количество различных факторов придется принимать в расчет. Однако сколь бы существенными ни были все эти соображения (а мы еще вернемся к ним в §2.6, Q8 и §3.5) они не имеют никакого отношения к тому, что называю «невычислимостью» я (и чего требует  $\mathcal{C}$ ). Под «невычислимостью» я подразумеваю принципиальную невозможность вычисления в том смысле, который мы очень скоро обсудим. Вычисления, которые просто выходят за рамки существующих или воображимых компьютеров, или имеющихся в нашем распоряжении вычислительных методов, формально всё равно остаются «вычислениями».

Читатель имеет полное право спросить: если ничего, что можно счесть «невычислимым», не обнаруживается ни в случайности, ни во влиянии окружения, ни в банальном несоответствии уровня сложности феномена нашим техническим возможностям, то что вообще я имею в виду, говоря «чего требует  $\mathcal{C}$ »? В общем случае, это некий вид математически точной активности, невычислимость которой можно доказать. Насколько нам на данный момент известно, при описании физического поведения в подобной математической активности необходимости не возникает. Тем не менее, логически она возможна. Более того, она представляет собой нечто большее, нежели просто логическую возможность. Согласно приводимой далее в книге аргументации, возможность активности подобного общего характера прямо подразумевается физическими законами, несмотря на то, что ни с чем подобным в известной физике мы еще не встречались. Некоторые примеры такой математической активности замечательно просты, поэтому представляется вполне уместным проиллюстрировать с их помощью то, о чем я здесь говорю.

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

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

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

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

Вот еще один пример:

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

Решением первой системы является, в частности, следующее:

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

тогда как вторая система вообще не имеет решения (судя по первому уравнению, число  $y$  должно быть четным, судя по второму уравнению, число  $z$  также должно быть четным, однако это противоречит третьему уравнению, причем при любом  $w$ , поскольку значение разности  $w^2 - w - 2$  это

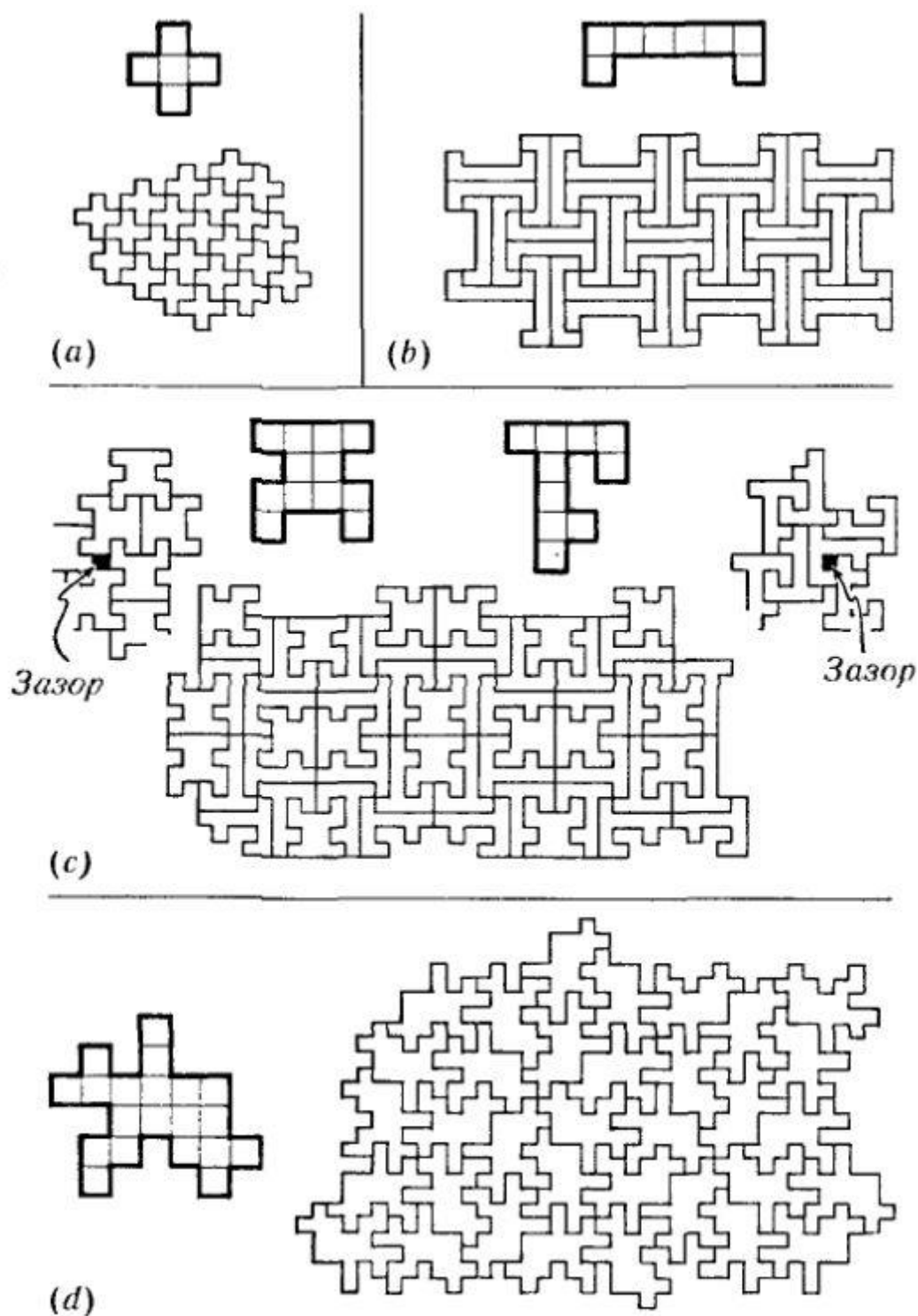
---

решает задачу типа «выполнить работу одного устройства на другом устройстве»; вместо этого он обсуждает совсем другую задачу: смоделировать (каким-то абстрактным и для меня, профессионального программиста, совершенно диким способом) мозг и его окружение на компьютерах при помощи методов «математического моделирования» или чего-то там подобного. (Ну и соответственно: если он докажет, что такое моделирование невозможно – а я охотно верю, что это так, ибо столь дикая идея мне без Пенроуза и в голову не могла бы придти! – то доказательство это будет относиться ко второй задаче (моделированию), и НИЧЕГО не скажет о первой задаче – выполнении работы мозга на другом устройстве. Вот, как, оказывается, всё просто!).

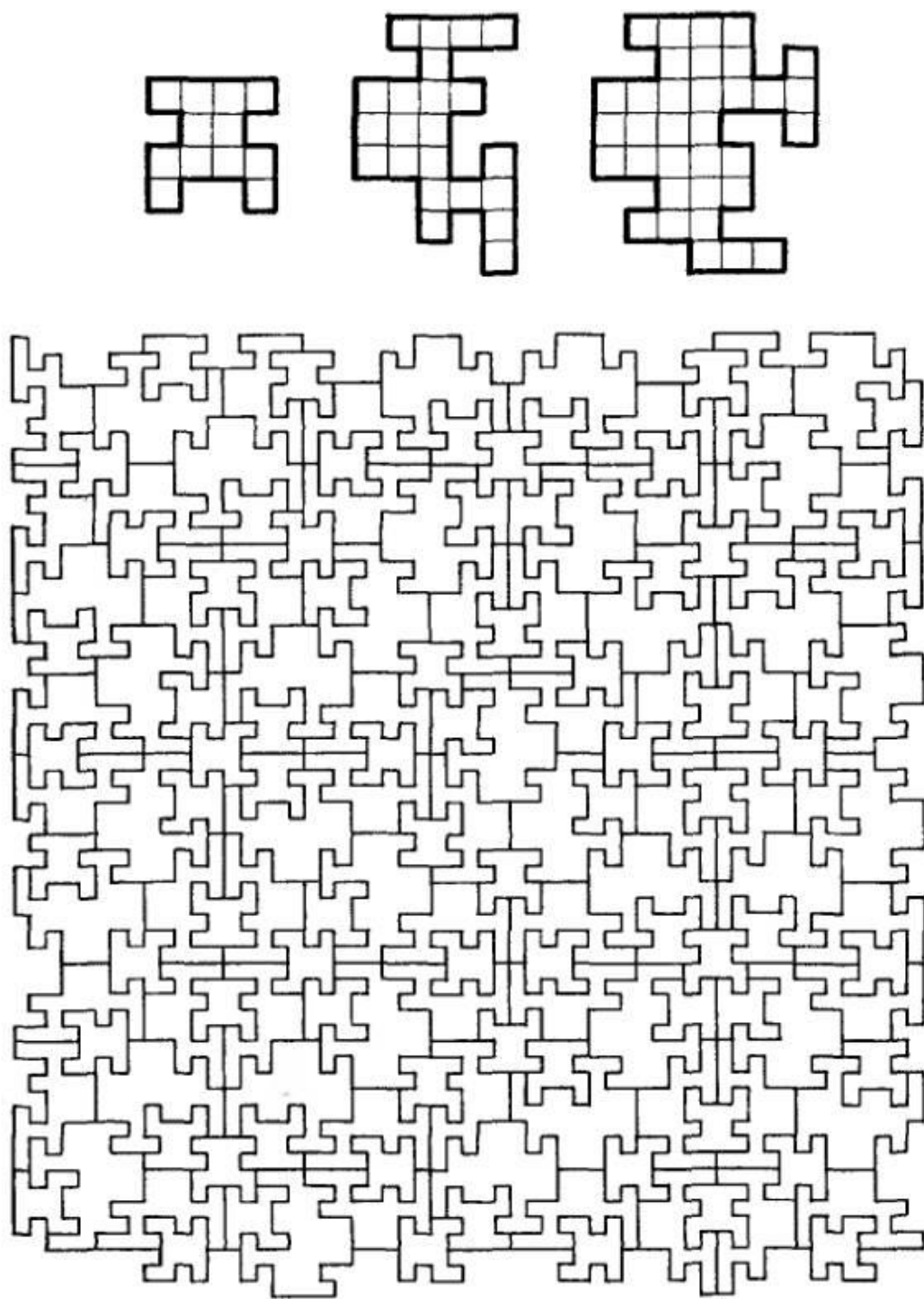
всегда четное число, а число 3 нечетно). Задача, поставленная Гильбертом, заключалась в отыскании математической процедуры (или алгоритма), позволяющей определить, какие системы диофантовых уравнений имеют решения (наш первый пример), а какие нет (второй пример). Вспомним (см. §1.5), что алгоритм – это всего лишь вычислительная процедура, действие некоторой машины Тьюринга. Таким образом, решением десятой проблемы Гильберта является некая вычислительная процедура, позволяющая определить, когда система диофантовых уравнений имеет решение.

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

Еще одним примером класса вполне структурированных математических задач, не имеющих алгоритмического решения, является задача о замощении. Она формулируется следующим образом: дан набор многоугольников, требуется определить, покрывают ли они плоскость; иными словами, возможно ли покрыть всю евклидову плоскость только этими многоугольниками без зазоров и наложений? В 1966 году американский математик Роберт Бергер показал (причем эффективно), что эта задача вычислительными средствами неразрешима. В основу его доводов легло обобщение одной из работ американского математика китайского происхождения Хао Вана, опубликованной в 1961 году (см. [176]). Надо сказать, что в моей формулировке задача оказывается несколько более громоздкой, чем хотелось бы, так как многоугольные плитки описываются в общем случае с помощью вещественных чисел (чисел, выражаемых в виде бесконечных десятичных дробей), тогда как обычные алгоритмы способны оперировать только целыми числами. От этого неудобства можно избавиться, если в качестве рассматриваемых многоугольников выбрать плитки, состоящие из нескольких квадратов, примыкающих один к другому сторонами. Такие плитки называются полиомино (см. [161]; [136], глава 13; [222]). На рис. 1.2 показаны некоторые плитки полиомино и примеры замощений ими плоскости. (Другие примеры замощений плоскости наборами плиток см. в НРК, с. 133–137, рис. 4.6–4.12.)



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



**Рис. 1.3.** Набор из трех полимино, покрывающий плоскость аperiodически (получен из набора Роберта Аммана).

Любопытно, что вычислительная неразрешимость задачи о замощении связана с существованием наборов полимино, называемых аperiodическими, такие наборы покрывают плоскость исключительно аperiodически (т.е. так, что никакой участок законченного узора нигде не повторяется, независимо от площади покрытой плиткой плоскости). На рис. 1.3 представлен аperiodический набор из трех полимино (полученный из набора, обнаруженного Робертом Амманом в 1977 году; см. [176], рис. 10.4.11–10.4.13 на с. 555–556).

Математические доказательства неразрешимости с помощью вычислительных методов десятой проблемы Гильберта и задачи о замощении весьма сложны, и я, разумеется, не стану и

пытаться приводить их здесь.<sup>57</sup> Центральное место в каждом из этих доказательств отводится, в сущности, тому, чтобы показать, каким образом можно запрограммировать машину Тьюринга на решение задачи о диофантовых уравнениях или задачи о замощении. В результате всё сводится к вопросу, который Тьюринг рассматривал еще в своем первоначальном исследовании: к вычислительной неразрешимости проблемы остановки – проблемы определения ситуаций, в которых работа машины Тьюринга не может завершиться. В §2.3 мы приведем несколько примеров явных вычислительных процедур, которые принципиально не могут завершиться, а в §2.5 будет представлено достаточно простое доказательство – основанное, по большей части, на оригинальном доказательстве Тьюринга – которое, помимо прочего, показывает, что проблема остановки действительно неразрешима вычислительными методами. (Что же касается следствий из того самого «прочего», ради которого, собственно, и затевалось упомянутое доказательство, то на них, в сущности, построены рассуждения всей первой части книги.)

$$\begin{aligned}
 S_0 &= \{ \}, & S_1 &= \{ \square \}, & S_2 &= \{ \begin{array}{|c|c|} \hline \square & \square \\ \hline \end{array} \}, & S_3 &= \{ \begin{array}{|c|c|} \hline \square & \square \\ \hline \square & \square \end{array} \}, \\
 S_4 &= \{ \begin{array}{|c|c|} \hline \square & \square \\ \hline \square & \square \end{array} \}, & S_5 &= \{ \begin{array}{|c|} \hline \square \\ \hline \square \\ \hline \end{array} \}, & S_6 &= \{ \begin{array}{|c|c|} \hline \square & \square \\ \hline \square & \square \end{array} \} \dots, \\
 S_{278} &= \{ \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \square & \square & \square \end{array} \}, \dots, & S_{975032} &= \{ \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \square & \square & \square \end{array}, \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \square & \square & \square \end{array}, \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \square & \square & \square \end{array} \}, \dots
 \end{aligned}$$

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

Каким же образом можно применить такой класс задач, как задачи о диофантовых уравнениях или задачи о замощении, к созданию «игрушечной» вселенной, которая, будучи детерминированной, является, тем не менее, невычислимой? Допустим, что в нашей модели вселенной течет дискретное время, параметризованное натуральными (т.е. целыми неотрицательными) числами  $0, 1, 2, 3, 4, \dots$ . Предположим, что в некий момент времени  $n$  состояние вселенной точно определяется одной задачей из рассматриваемого класса, скажем, набором полимино. Необходимо установить два вполне определенных правила относительно того, какой из наборов полимино будет представлять состояние вселенной в момент времени  $n+1$  при заданном наборе полимино для состояния вселенной в момент времени  $n$ , причем первое из этих

<sup>57</sup> В действительности Роберт Бергер доказал, что общего алгоритмического решения не имеет лишь задача о замощении плоскости плитками Вана. Плитки Вана (названные так в честь математика Хао Вана) представляют собой единичные квадраты с окрашенными краями; при замощении цвета соседних плиток должны совпадать, сами же плитки при этом нельзя ни вращать, ни переворачивать. Впрочем, для любого набора плиток Вана несложно составить такой набор полимино, которым можно будет замостить плоскость тогда и только тогда, когда ее можно замостить соответствующим набором плиток Вана. Таким образом, неразрешимость вычислительными методами задачи о замощении плоскости набором полимино непосредственно следует из неразрешимости задачи о замощении плоскости набором плиток Вана. В связи с задачей о замощении плоскости полимино следует отметить, что если каким-либо набором полимино не удается замостить плоскость, то этот факт вполне возможно установить вычислительным путем (точно так же, как мы можем предсказать остановку машины Тьюринга или убедиться в наличии решения у системы диофантовых уравнений), нужно лишь попытаться замостить плитками данного набора квадратную область размера  $n \times n$  (последовательно увеличивая значение  $n$ ); замостить всю плоскость не удастся уже при некотором конечном значении  $n$ . Алгоритмическим путем невозможно установить как раз те случаи, когда данным набором плиток можно-таки замостить плоскость.

правил применяется в том случае, если полиоминно покрывают всю плоскость без зазоров и наложений, а второе – если это не так. То, как именно будут выглядеть подобные правила, не имеет в данном случае особого значения. Можно составить список  $S_0, S_1, S_2, S_3, S_4, S_5, \dots$  всех возможных наборов полиоминно таким образом, чтобы наборы, содержащие в общей сложности четное число квадратов, имели бы четные индексы  $S_0, S_2, S_4, S_6, \dots$  а наборы с нечетным количеством квадратов – нечетные индексы  $S_1, S_3, S_5, S_7, \dots$  (Составление такого списка не представляет особой сложности; нужно лишь подобрать соответствующую вычислительную процедуру.) Итак, «динамическая эволюция» нашей игрушечной вселенной задается теперь следующим условием:

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

Поведение такой вселенной полностью детерминировано,<sup>58</sup> однако поскольку в нашем распоряжении нет общей вычислительной процедуры, позволяющей установить, какой из наборов полиоминно  $S_n$  покрывает плоскость (причем это верно и тогда, когда общее число квадратов постоянно, независимо от того, четное оно или нет), то невозможно и численное моделирование ее реального развития. (См. рис. 1.4.)

Безусловно, такую схему нельзя воспринимать хоть сколько-нибудь всерьез – она ни в коем случае не моделирует реальную вселенную, в которой все мы живем. Эта схема приводится здесь (как, собственно, и в НРК, с. 170) для иллюстрации того часто недооцениваемого факта, что между детерминизмом и вычислимостью существует вполне определенная разница. Некоторые полностью детерминированные модели вселенной с четкими законами эволюции невозможно реализовать вычислительными средствами. Вообще говоря, как мы убедимся в §7.9, только что рассмотренные мною весьма специфические модели не совсем отвечают реальным требованиям точки зрения  $\mathcal{C}$ . Что же касается тех феноменов, которые отвечают-таки этим самым реальным требованиям, и некоторых связанных с упомянутыми феноменами поразительных физических возможностях, то о них мы поговорим в §7.10.

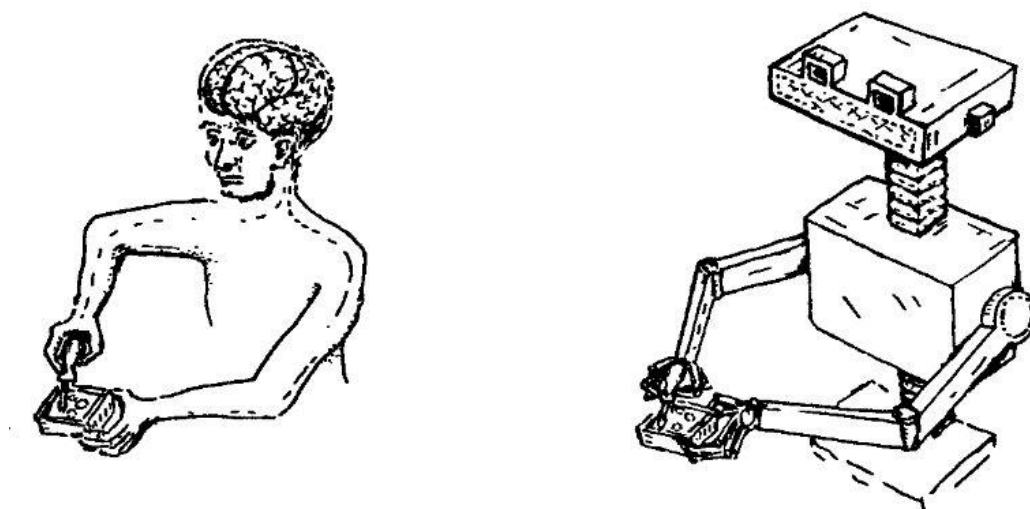
### §1.10. Завтрашний день

Так какого же будущего для этой планеты нам следует ожидать согласно точкам зрения  $\mathcal{A}$ ,  $\mathcal{B}$ ,  $\mathcal{C}$  и  $\mathcal{D}$ ? Если верить  $\mathcal{A}$ , то настанет время, когда соответствующим образом запрограммированные суперкомпьютеры догонят – а затем и перегонят человека во всех его интеллектуальных достижениях. Конечно же, сторонники  $\mathcal{A}$  придерживаются различных взглядов относительно необходимого для этого времени. Некоторые вполне разумно полагают, что пройдет еще много столетий, прежде чем компьютеры достигнут уровня человека, принимая во внимание крайнюю скудость современного понимания реально выполняемых мозгом вычислений (так они говорят), обуславливающих ту тонкость поведения, какую, несомненно, демонстрирует человек, – тонкость, без которой, конечно же, нельзя говорить о каком бы то ни было «пробуждении сознания». Другие утверждают, что времени понадобится значительно меньше. В частности, Ханс Моравек в своей книге «Дети разума» [267] приводит вполне аргументированное доказательство (основанное на непрерывно ускоряющемся развитии компьютерных технологий за последние пятьдесят лет и на своей оценке той доли от всего объема функциональной активности мозга, которая на сегодняшний день уже успешно моделируется численными методами) в поддержку своего утверждения, будто уровень «эквивалентности человеку» будет

---

<sup>58</sup> В.Э.: Оно детерминировано лишь в некоторой абстрактной, идеальной модели. В реальном мире такая вселенная не может существовать (если мы принимаем те постулаты, на которых основывается мое мировоззрение в целом и Веданская теория в частности), потому, что здесь в причинно-следственной цепочке принимается в качестве действительной причины решение одной программы, притом не программы существующей, а программы несуществующей – которую нельзя написать, так как для нее нет алгоритма. Ведь фактор «покрывает плоскость» или «не покрывает плоскость» не существует реально в материальном мире. Реально существует только задачка, которую мы сами себе поставили: «попытаться вот такими вот фигурками покрыть всё плоское пространство». Для решения этой задачки нам нужно было составить (создать самопрограммированием) мозговую программу ее решения. Но мы такую программу создать не смогли. И вот – исход этой (так и не созданной) мозговой программы становится реальной причиной дальнейших событий во Вселенной!

преодолен уже к 2030 году. (Кое-кто утверждает, что это время будет еще короче,<sup>59</sup> а кто-то даже уверен, что предсказанная дата достижения эквивалентности человеку уже осталась в прошлом!) Однако чтобы читатель не очень пугался того, что менее чем через сорок (или около того) лет компьютеры во всем его превзойдут, горькая пилюля подслащена одной радужной надеждой (подаваемой под видом гарантированного обещания): все мы сможем тогда перенести свои «ментальные программы» в сверкающие металлические (или пластиковые) корпуса роботов (конкретную модель, разумеется, каждый выберет себе сам), чем и обеспечим себе что-то вроде бессмертия [267, 268].



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

А вот сторонники точки зрения  $\mathcal{B}$  подобным оптимизмом похвастаться не могут. Они вполне согласны с приверженцами  $\mathcal{A}$  относительно перспектив развития интеллектуальных способностей компьютеров – с той лишь оговоркой, что речь при этом идет исключительно о внешних проявлениях этих самых способностей. Для управления роботом необходимо и достаточно располагать адекватной моделью деятельности человеческого мозга, больше ничего не требуется (рис. 1.5). Согласно  $\mathcal{B}$ , вопрос о том, способно ли подобное моделирование вызвать осмысленное осознание, не имеет никакого отношения к реальному поведению робота. На достижение необходимого для такого моделирования технологического уровня может уйти как несколько веков, так и менее сорока лет. Однако, как уверяют сторонники  $\mathcal{B}$ , рано или поздно, а это все-таки произойдет. Тогда же компьютеры достигнут уровня «эквивалентности человеку», а затем, как можно ожидать, и уверенно превзойдут его, оставив без внимания все потуги нашего относительно слабого мозга хоть немного этот уровень приподнять. Причем возможности «подключения» к управляемым роботам у нас в этом случае не будет, и, похоже, придется примириться с тем, что нашей планетой, в конечном итоге, будут править абсолютно бесчувственные машины! Мне представляется, что из всех точек зрения  $\mathcal{A}$ ,  $\mathcal{B}$ ,  $\mathcal{C}$  и  $\mathcal{D}$  именно  $\mathcal{B}$  предлагает самый пессимистичный взгляд на будущее нашей планеты – вопреки, казалось бы, тому факту, что именно она лучше всего соотносится с так называемым «здравым смыслом».

Если же верить  $\mathcal{C}$  или  $\mathcal{D}$ , то можно ожидать, что компьютеры навсегда сохранят подчиненное по отношению к человеку положение – какими бы быстрыми, мощными или алгоритмически совершенными они ни стали. При этом точка зрения  $\mathcal{C}$  не отрицает возможности будущих научных разработок, которые могут привести к созданию неких устройств, принцип действия которых не будет иметь ничего общего с компьютерами в их сегодняшнем понимании, а будет основан на той самой невычислимой физической активности, которая, согласно  $\mathcal{C}$ , обуславливает наше собственное сознательное мышление, – устройств, которые окажутся

<sup>59</sup> О некоторых чересчур оптимистичных прогнозах относительно ИИ можно прочесть в [124].

способны вместить в себя реальные разум и сознание.<sup>60</sup> Быть может, в конечном итоге именно такие устройства, а вовсе не те машины, которые мы называем «компьютерами», и превзойдут человека в интеллектуальном отношении. Что ж, не исключено; однако подобные умозрительные прогнозы представляются мне в настоящий момент крайне преждевременными, поскольку мы практически не обладаем необходимыми для таких исследований научными познаниями, не говоря уже о каких бы то ни было технологических решениях. К этому вопросу мы еще вернемся во второй части книги (§8.1).

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

С некоторых пор умы теоретиков от юриспруденции начал занимать один вопрос, имеющий самое непосредственное отношение к теме нашего разговора, но в некотором смысле более практический.<sup>61</sup> Суть его заключается в следующем: не предстоит ли нам в не столь отдаленном будущем задуматься над тем, обладают ли компьютеры законными правами и несут ли они ответственность за свои действия. В самом деле, если со временем компьютеры смогут достичь уровня человека (а то и превзойти его) в самых разных областях деятельности, то подобные вопросы неминуемо должны приобрести определенную значимость. Если придерживаться точки зрения *A*, то следует, очевидно, признать, что компьютеры (или управляемые компьютером роботы) должны потенциально и обладать правами, и нести ответственность.<sup>62</sup> Ибо, согласно этой точке зрения, между человеком и роботом достаточно высокого уровня сложности нет существенной разницы, за исключением такой «мелочи», как различие в материальном строении. Однако приверженцам точки зрения *B* ситуация представляется несколько более запутанной. Разумно утверждать, что вопрос о правах или ответственности уместен для созданий, наделенных способностью чувствовать, т.е. испытывать определенные, подлинно душевные «ощущения» – такие, как страдание, гнев, мстительность, злоба, вера (религиозная и общечеловеческая), желание, сомнение, понимание или страсть. Согласно *B* управляемый компьютером робот не обладает такой способностью, вследствие чего, на мой взгляд, не может ни обладать правами, ни нести ответственность. С другой стороны, если верить *B*, не существует эффективного способа определить, что упомянутая способность у робота действительно отсутствует, поэтому если роботы смогут достаточно правдоподобно имитировать поведение человека, то человек может оказаться в весьма затруднительном положении.

Подобного затруднения, по всей видимости, не возникнет у сторонников точки зрения *C* (а также, возможно, *D*), поскольку, согласно этим точкам зрения, компьютеры не в состоянии убедительно демонстрировать душевные переживания и, уж конечно же, ничего похожего не чувствуют и чувствовать никогда не будут. Соответственно, компьютеры не могут ни обладать правами, ни нести ответственность. Лично мне такая точка зрения представляется весьма разумной. Вообще в этой книге я выступаю как серьезный противник позиций *A* и *B*. Согласившись с моими аргументами, юристы, безусловно, существенно упростят себе жизнь: как таковые компьютеры или управляемые компьютерами роботы ни при каких обстоятельствах не обладают правами и не несут ответственности.<sup>63</sup> Нельзя обвинить компьютеры в каких бы то ни было неприятностях или недоразумениях – виновен всегда человек!

Следует, однако, понимать, что вышеприведенные аргументы могут и не относиться к всевозможным гипотетическим «устройствам», подобным упомянутым выше – тем, что смогут в

---

<sup>60</sup> В.Э.: Таковы уже современные компьютеры (точнее: не те, которые производятся промышленностью для офисов, а те, которые уже при нынешнем уровне технологий могли бы создаваться специально для искусственного интеллекта). Но для создания ИИ требуется разработка специальных компьютеров и разработка специальной программатуры, что очень дорого и вряд ли будет кем-нибудь финансироваться. (Не говоря уже о том сопротивлении – подобном борьбе против клонирования людей, – которое такой проект встретил бы с разных сторон).

<sup>61</sup> Своим знакомством с этими вопросами я обязан очень многим людям, среди которых хочу особо поблагодарить Ли Левингера. Замечательное исследование связи современной физики и вычислительных методов с проблемами человеческого поведения можно найти в книге [200].

<sup>62</sup> В.Э.: Противоположное будет означать расизм. Правомерен ли расизм – это другой вопрос.

<sup>63</sup> В.Э.: То, что встроенный в компьютер реальный разум должен обладать одинаковыми с человеком правами и ответственностью – это очевидно. Но проблема здесь кроется совсем в другом: может ли обладать правами и нести ответственность за свои действия ЧЕЛОВЕК? (И я вообще-то склоняюсь к ответу: НЕТ).

конечном итоге воплотить в себе принципы новой, невычислительной физики. Но, поскольку перспектива появления таких устройств – если их вообще удастся создать – весьма туманна, возникновения связанных с ними юридических проблем в ближайшем будущем ожидать не приходится.

Проблема «ответственности» поднимает глубокие философские вопросы, связанные с основными факторами, обуславливающими наше поведение. Можно вполне обоснованно утверждать, что каждое наше действие так или иначе определяется наследственностью и окружением, а то и всевозможными случайностями, непрерывно влияющими на нашу жизнь. Но ведь ни одно из этих воздействий никак не зависит лично от нас, почему же мы должны нести за них ответственность? Является ли понятие «ответственности» лишь терминологической условностью, или дело в чем-то еще? Возможно, и впрямь существует некая «самость» – нечто, стоящее «выше» уровня подобных влияний и определяющее, в конечном счете, наши действия? В юридическом смысле понятие «ответственности» явно подразумевает, что внутри каждого из нас и в самом деле существует своего рода независимая «самость», наделенная своей собственной ответственностью – и, по определению, правами, – причем ее проявления нельзя объяснить ни наследственностью, ни окружением, ни случайностью.<sup>64</sup> Если же присутствие в нашей речи такой независимой «самости» не просто языковая условность, то в современных физических представлениях недостает чего-то весьма существенного. Открытие этого недостающего ингредиента, несомненно, многое изменит в нашем научном мировоззрении.

Хотя книга, которую вы держите в руках, и не дает исчерпывающего ответа на эти серьезные вопросы, она, как я полагаю, может чуть приоткрыть дверь, отделяющую нас от них – не больше, но и не меньше. Вы не найдете здесь неопровержимых доказательств неперенного существования такой «самости», проявления которой нельзя объяснить никакой внешней причиной, вам лишь предложат несколько шире взглянуть на самую природу возможных «причин». «Причина» может оказаться невычислимой – на практике или в принципе. Я намерен показать, что если упомянутая «причина» так или иначе порождается нашими сознательными действиями, то она должна быть весьма тонкой, безусловно невычислимой и не имеющей ничего общего ни с хаосом, ни с прочими чисто случайными воздействиями. Сможет ли такая концепция «причины» приблизить нас к пониманию истинной сущности свободы воли (или иллюзорности такой свободы)<sup>65</sup> – вопрос будущего.

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

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

Хотя я не вижу особой необходимости пытаться дать непременно полные определения, некоторые комментарии относительно моей собственной терминологии представляются всё же уместными. Я часто с некоторым замешательством обнаруживаю, что употребление всех этих слов, столь очевидное для меня, не совпадает с тем, что полагают естественным другие.

---

<sup>64</sup> В.Э.: Наш мир полностью детерминирован, и наша «самость», естественно, является лишь одним звеном в одной цепочке причинно-следственных отношений Вселенной. В этом смысле каждый из нас – результат «наследственности» и «окружения» (хотя когда люди об этом говорят, они обычно представляют себе всё это не совсем точно и правильно). Но в то же время эта «самость» и ответственна за то, какая она есть. Здесь нет никакого противоречия: оно надуманно и основано на непонимании положения вещей. А положение здесь точно такое же, как и с системой программ (и мы и есть система программ!). Если программы работают неправильно (с точки зрения каких-то критериев, например: Уголовного кодекса), то программы надо либо исправлять, либо уничтожать.

<sup>65</sup> В.Э.: Вопрос о «свободе воли» похож на вопрос об ответственности. Разумеется, наше поведение стопроцентно детерминировано внутримозговыми процессами. Но нет другого «Я», кроме этих процессов, поэтому «мое решение» совпадает с «решением этих процессов» – это одно и то же. И человек обладает «свободой воли», несмотря на полную детерминированность мира. И никакого противоречия здесь нет. (У человека не было бы «свободы воли» только в том случае, если бы его «Я» было одной вещью, а мозговые процессы – другой вещью, и, вот, тогда эта вторая вещь принуждала бы первую к чему-то).

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

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

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

Лично мне такая терминология кажется совершенно естественной, однако многие поборники ИИ (во всяком случае те из них, кто не поддерживает точку зрения *A*) станут решительно отрицать всякую свою причастность к попыткам реализации искусственного «осознания», хотя конечной их целью является, судя по названию, не что иное, как искусственный «интеллект»<sup>66</sup>. Они, пожалуй, оправдаются тем, что они (в полном согласии с *B*)

<sup>66</sup> В.Э.: Вообще в английском языке с этими терминами дело обстоит несколько иначе, чем в русском. То, что по-русски обычно переводится как «искусственный интеллект», по-английски называется «Artificial intelligence», а не «Artificial intellect», и первый термин означает не столько полный интеллект (разум), сколько отдельные умственные способности. Термин этот вышел на мировую арену 31 августа 1955 года с разосланным Джоном Маккарти и Мервином Минским Приглашением разным ученым на Дартмутскую конференцию по «искусственной интеллигентности». По этому документу (а он доступен в Интернете) особенно ярко видно, что речь не шла о создании полного искусственного интеллекта, а только о воспроизведении на компьютерах отдельных черт «разумного» поведения. Мне кажется, что русские переводчики не всегда учитывают эту тонкость.

всего лишь моделируют интеллект – такая модель не требует действительного понимания или осознания, – а вовсе не пытаются создать то, что я называю подлинным интеллектом. Вероятно, они будут уверять вас, что не видят никакой разницы между подлинным интеллектом и его моделью, что вполне отвечает точке зрения <sup>А</sup>. В своих дальнейших рассуждениях я, в частности, намерен показать, что некоторые аспекты «подлинного понимания» действительно невозможно воссоздать путем каких бы то ни было вычислений. Следовательно, должно существовать и различие между подлинным интеллектом и любой попыткой его достоверного численного моделирования.

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

а) «интеллект» требует «понимания»

и

б) «понимание» требует «осознания».

Осознание я воспринимаю как один из аспектов – пассивный – феномена сознания. У сознания имеется и активный аспект, а именно – свободная воля. Полного определения слова «сознание» здесь также не дается (и, уж конечно же, не мне определять, что есть «свободная воля»), хотя мои аргументы имеют целью окончательное объяснение феномена сознания в научных, но невычислительных терминах – как того требует точка зрения <sup>С</sup>. Не претендую я и на то, что мне удалось преодолеть хоть сколько-нибудь значительное расстояние на пути к этой цели, однако надеюсь, что представленная в этой книге (равно как и в НРК) аргументация расставит вдоль этого пути несколько полезных указателей для идущих следом – а может, станет и чем-то большим. Мне кажется, что, пытаясь на данном этапе дать слишком точное определение термину «сознание», мы рискуем упустить ту самую концепцию, какую хотим изловить. Поэтому вместо поспешного и наверняка неадекватного определения я приведу лишь несколько комментариев описательного характера относительно моего собственного употребления термина «сознание». В остальном же нам придется положиться на интуитивное понимание смысла этого термина.

Всё это вовсе не означает, что я полагаю, будто мы действительно «интуитивно знаем», чем на самом деле «является» сознание; я лишь хочу сказать, что такое понятие существует, а мы, по мере сил, пытаемся его постичь – причем за понятием стоит некий реально существующий феномен,<sup>67</sup> который допускает научное описание и играет в физическом мире как пассивную, так и активную роль. Некоторые, судя по всему, полагают, что данная концепция слишком туманна, чтобы заслуживать серьезного изучения. Однако при этом те же люди<sup>68</sup> часто и с удовольствием рассуждают о «разуме», полагая, очевидно, что это понятие определено гораздо точнее. Общепринятое употребление слова «разум» предполагает разделение этого самого разума (возможное или реальное) на так называемые «сознательную» и «бессознательную» составляющие. На мой взгляд, концепция бессознательного разума представляется еще более невразумительной, нежели концепция разума сознательного. Я и сам нередко пользуюсь словом «разум», однако не пытаюсь при этом дать его точное определение. В нашей

<sup>67</sup> В.Э.: В бытовом (а, значит, и в теперешнем «научном») языке под словами с основой «..сознан..» понимаются различные вещи, но если из них отобрать самое существенное, самое центральное (как это и пытается сделать Пенроуз), то этот «феномен» можно определить так: это способность одних программ компьютера анализировать «со стороны» другие программы, предварительно оценивать их потенциальные последствия, помнить, оценивать и принимать во внимание в дальнейшем самопрограммировании реальные результаты их выполнения.

<sup>68</sup> Сломен [344], например, пеняет мне на то, что в НРК я слишком часто прибегаю к такому неопределенному термину, как «сознание», в то время как сам он весьма свободно оперирует еще более неопределенным (на мой взгляд) термином «разум»!

последующей дискуссии (достаточно строгой, надеюсь) концепция «разума» – за исключением той ее части, что уже нашла свое воплощение в термине «сознание», – не будет играть центральной роли.

Что же я имею в виду, говоря о сознании? Как уже отмечалось ранее, сознание обладает активным и пассивным аспектами, однако различие между ними далеко не всегда чётко определено. Восприятие, скажем, красного цвета требует несомненно пассивного сознания, равно как и ощущение боли либо восхищение музыкальным произведением. Активное же сознание участвует в сознательных действиях – таких, например, как подъем с кровати или, напротив, намеренное решение воздержаться от какой-либо энергичной деятельности. При воссоздании в памяти каких-то прошедших событий оказываются задействованы как пассивный, так и активный аспекты сознания. Составление плана будущих действий также обычно требует участия сознания – и активного, и пассивного; и, надо полагать, какое-никакое сознание необходимо для умственной деятельности, которую общепринято описывать словом «понимание». Более того, мы остаемся, в определенном смысле, в сознании (пассивный аспект), даже когда спим, если при этом нам снится сон (в процессе же пробуждения может принимать участие и активный аспект сознания).

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

В первой части книги меня будет занимать, главным образом, вопрос о том, чего можно достичь, используя свойство мышления, известное как «понимание». Хотя я не даю здесь определения термину «понимание», надеюсь всё же прояснить его смысл в достаточной мере для того, чтобы убедить читателя в том, что обозначаемое этим термином свойство – чем бы оно ни оказалось – и в самом деле должно быть неотъемлемой частью мыслительной деятельности, которая необходима, скажем, для признания справедливости рассуждений, составляющих §2.5. Я намерен показать, что восприятие этих рассуждений должно быть связано с какими-то принципиально невычислимыми процессами. Мое доказательство не затрагивает столь непосредственно другие свойства мыслительной деятельности («интеллект», «осознание», «сознание» или «разум»), однако оно имеет определенное отношение и к этим концепциям, поскольку, в соответствии с той терминологией «от здравого смысла», о которой я упоминал выше, осознание непременно должно быть существенным компонентом понимания, а понимание – являться неотъемлемой частью любого подлинного интеллекта.<sup>69</sup>

### §1.13. Доказательство Джона Серла

Прежде чем представить свое собственное рассуждение, хотелось бы вкратце упомянуть о совсем иной линии доказательства – знаменитой «китайской комнате» философа Джона Серла<sup>70</sup> – главным образом для того, чтобы подчеркнуть существенное отличие от нее моего доказательства как по общему характеру, так и по базовым концепциям. Доказательство Серла тоже связано с проблемой «понимания» и имеет целью выяснить, можно ли утверждать, что функционирование достаточно сложного компьютера реализует это свойство мышления. Я не буду повторять здесь рассуждение Серла во всех подробностях, а лишь кратко обозначу его суть.

Дана некая компьютерная программа, которая демонстрирует имитацию «понимания», отвечая на вопросы о какой-то рассказанной ей предварительно истории, причем все вопросы и

<sup>69</sup> В.Э.: Ну, в общем ничего существенного по определению этих понятий Пенроуз не сказал...

<sup>70</sup> См. [340], [341].

ответы даются на китайском языке. Далее Серл рассматривает не владеющего китайским языком человека, который старательно воспроизводит все до единой вычислительные операции, выполняемые в процессе имитации компьютером. При этом когда вычисления выполняет компьютер, получаемые на его выходе данные создают некоторую видимость понимания; когда же все необходимые вычисления посредством соответствующих манипуляций воспроизводит человек, какого-либо понимания в действительности не возникает. На этом основании Серл утверждает, что понимание как свойство мышления не может сводиться исключительно к вычислениям – хотя человек (не знающий китайского) и воспроизводит каждую вычислительную операцию, выполняемую компьютером, он всё же совершенно не понимает смысла рассказанной истории. Серл допускает, что возможно осуществить моделирование получаемых на выходе результатов понимания (в полном соответствии с точкой зрения  $\mathcal{B}$ ), поскольку он полагает, что это вполне достижимо посредством компьютерного моделирования всей физической активности мозга (чем бы он при этом ни занимался) в тот момент, когда его владелец вдруг что-либо понимает. Однако главный вывод из «китайской комнаты» Джона Серла заключается в том, что сама по себе модель в принципе не способна действительно «ощутить» понимание. То есть для любой компьютерной модели подлинное понимание остается, в сущности, недостижимым.<sup>71</sup>

Доказательство Серла направлено против точки зрения  $\mathcal{A}$  (согласно которой любая «модель» понимания эквивалентна «подлинному» пониманию) и, по замыслу автора, в поддержку точки зрения  $\mathcal{B}$  (хотя в той же мере оно поддерживает и  $\mathcal{C}$  или  $\mathcal{D}$ ). Оно имеет дело с пассивным, обращенным внутрь, или субъективным аспектами понимания, однако при этом не отрицает возможности моделирования понимания в его активном, обращенном наружу, или объективном аспектах. Сам Серл однажды заявил: «Несомненно, мозг – это цифровой компьютер. Раз кругом одни цифровые компьютеры, значит, и мозг должен быть одним из них»<sup>72</sup>. Отсюда можно заключить, что Серл готов принять возможность полного моделирования работы обладающего сознанием мозга в процессе «понимания», результатом которого оказалась бы полная тождественность внешних проявлений модели и внешних проявлений действительно мыслящего человеческого существа, что соответствует точке зрения  $\mathcal{B}$ . Мое же исследование призвано показать, что одними лишь внешними проявлениями «понимание» отнюдь не ограничивается, в связи с чем я утверждаю, что невозможно построить достоверную компьютерную модель даже внешних проявлений понимания.<sup>73</sup> Я не привожу здесь аргументацию Серла в подробностях, поскольку точку зрения  $\mathcal{C}$  она напрямую не поддерживает (а целью всех наших дискуссий здесь является как раз поддержка  $\mathcal{C}$  и ничто иное). Тем не менее, следует отметить, что концепция «китайской комнаты» предоставляет, на мой взгляд, достаточно убедительный аргумент против  $\mathcal{A}$ ,<sup>74</sup> хотя я и не считаю этот аргумент решающим. Более подробное изложение и различные контраргументы представлены в [340], обсуждение – там же и в [203]; см. также [80] и [341]. Мою оценку можно найти в НРК, с. 17–23.

#### §1.14. Некоторые проблемы вычислительной модели

Прежде чем перейти к вопросам, отражающим специфические отличия точки зрения  $\mathcal{C}$  от  $\mathcal{A}$  и  $\mathcal{B}$ , рассмотрим некоторые другие трудности, с которыми непременно сталкивается любая попытка объяснить феномен сознания в соответствии с точкой зрения  $\mathcal{A}$ . Согласно  $\mathcal{A}$ , для возникновения осознания необходимо лишь простое «выполнение» или воспроизведение надлежащих алгоритмов. Что же это означает в действительности? Следует ли под «воспроизведением» понимать, что в соответствии с последовательными шагами алгоритма должны перемещаться с места на место некие физические материальные объекты? Предположим,

<sup>71</sup> В.Э.: Если компьютер имитирует разум, то понимание (и осознание) ему недоступны; если же в компьютере реализован разум, то он и понимает, и осознает. «Китайскую комнату» Серла я детально разобрал в мае 2000 года в своем ответе профессору Тамбергу по-латышски, а теперь готовлю перевод этих материалов на русский язык, поэтому здесь не буду повторяться, тем более, что это действительно очень обширное рассуждение.

<sup>72</sup> См. статью Серла [340] (ее также можно найти в сборнике [203], с. 372). Мне, правда, не совсем ясно, к какой точке зрения Серл склонился бы сейчас, к  $\mathcal{A}$  или всё же к  $\mathcal{B}$ .

<sup>73</sup> В.Э.: Модель построить невозможно, а реализовать можно.

<sup>74</sup> В.Э.: Против  $\mathcal{A}$  – да, но не против  $\mathcal{C}$ .

что эти последовательные шаги записываются строка за строкой в огромную книгу.<sup>75</sup> Являются ли «воспроизведением» действия, посредством которых осуществляется запись или печать этих строк? Достаточно ли одного лишь статического существования такой книги для осознания? А если просто водить пальцем от строчки к строчке – можно ли это считать «воспроизведением»? Или если водить пальцем по символам, набранным шрифтом Брайля? А если проецировать страницы книги одну за другой на экран? Является ли воспроизведением простое представление последовательных шагов алгоритма? С другой стороны, необходимо ли, чтобы кто-нибудь проверял, на самом ли деле каждая последующая линия надлежащим образом следует из предыдущей (в соответствии с правилами рассматриваемого алгоритма)? Последнее предположение способно, по крайней мере, разрешить все наши сомнения,<sup>76</sup> поскольку данный процесс должен, по всей видимости, обходиться без участия (сознательного) каких бы то ни было ассистентов. И всё же нет совершенно никакой ясности относительно того, какие именно физические действия следует считать действительными исполнителями алгоритма осознания. Быть может, подобные действия не требуются вовсе, и можно, не противореча точке зрения *A*, утверждать, что для возникновения «осознания» вполне достаточно одного лишь теоретического математического существования соответствующего алгоритма (см. § 1.17).

Как бы там ни было, можно предположить, что, даже согласно *A*, далеко не всякий сложный алгоритм может обусловить возникновение осознания (ощущения осознания). Наверное, для того, чтобы можно было считать состоявшимся сколько-нибудь заметное осознание, алгоритм, судя по всему, должен обладать некоторыми особыми свойствами – такими, например, как «высокоуровневая организация», «универсальность», «самоот-носимость», «алгоритмическая простота/сложность»<sup>77</sup> и тому подобными. Кроме того, донельзя скользким представляется вопрос о том, какие именно свойства алгоритма отвечают в этом случае за различные *qualia* (ощущения), формирующие осознание. Например, какое конкретно вычисление вызывает ощущение «красного»<sup>78</sup>? Какие вычисления дают ощущения «боли», «сладости», «гармоничности», «едкости» и т.д.? Сторонники *A* время от времени предпринимают попытки разобраться в подобного рода проблемах (см., например, [81]), однако пока что эти попытки выглядят весьма и весьма неубедительными.

Более того, любое четко определенное и достаточно простое алгоритмическое предположение (подобное всем тем, что до сих пор выдвигались в соответствующих исследованиях) об-

<sup>75</sup> Занимательное рассмотрение подобного предположения представлено в [202]; см. также НРК, с. 21–22.

<sup>76</sup> **В.Э.:** Все эти «сомнения» выглядят просто наивными для любого программиста. Дело ведь элементарное: возьмите любую компьютерную программу (хотя бы, например, *Word*), запишите ее инструкции азбукой Брайля, проведите пальцем по ним и посмотрите, какой будет эффект ☺. Программы, реализующие разум, – обыкновенные компьютерные программы; дело не в том, КАК их выполнять, а дело в том, КАКОЙ у них алгоритм, что они делают. (Пенроуз совершенно не может поставить и рассмотреть вопрос «Как работу мозга выполнить на другом устройстве?»; он рассматривает исключительно вопрос «Как мозг моделировать на компьютере?»).

<sup>77</sup> Суть понятия «алгоритмической сложности» доступным языком изложена в [46].

<sup>78</sup> **В.Э.:** Да, здесь есть некоторая философская проблема. Но мы (*ℰ*) ее решаем так. Если смотреть на человека объективно, «со стороны», то «ощущение красного» означает: 1) что эта система (человек) способна обнаруживать электромагнитные волны определенного (видимого) диапазона; 2) что она способна различать волны разной длины (отличить «красное» от других цветов); 3) что она (система) особое внимание обращает на волны «красного» диапазона (так как для предков системы эти волны означали либо съедобные фрукты на фоне зеленого леса, либо опасность – огонь пожара или пасть хищника –, либо готовые к совокуплению половые органы партнера), и это «особое внимание» системы проявляется как особая яркость этого цвета. Это то объективное, что стоит за «ощущением красного». Далее у нас есть два пути: (1) предположить, что это объективное и есть ВСЁ, что составляет «ощущение красного» и что этот объективный вид и наше субъективное ощущение – это просто взгляд извне и взгляд изнутри на одну и ту же вещь, и что они не отличаются по существу; и (2) предположить, что это объективно видимое явление – это еще не всё и что в добавок к нему существует еще что-то другое, какое-то еще «ощущение красного», которое возникает от объективного явления, но не совпадает с ним, а существует с ним рядом. Предположить-то мы так можем, но постулированная в предположении (2) дополнительная сущность не проявит себя ни в каких исследованиях, ни в каких экспериментах. Следовательно, все научные объяснения можно сделать БЕЗ нее, БЕЗ ее постулирования. И поэтому (на основании «лезвия Оккама») мы ее отмечаем. Остается предположение (1): что те названные выше три объективных пункта и наше субъективное «ощущение красного» – это одно и то же. Аналогично обстоит дело и с другими «ощущениями».

ладает одним существенным недостатком: этот алгоритм можно без особых усилий реализовать на современном электронном компьютере. А между тем, согласно утверждению автора такого предположения, реализация его алгоритма неизбежно вызывает реальное ощущение того или иного *qualium*.<sup>79</sup> Мне думается, что даже самому стойкому приверженцу точки зрения *A* будет сложно всерьез поверить, что такое вычисление – да и вообще любое вычисление, которое можно запустить на современном компьютере, работа которого основывается на современных представлениях об ИИ, – может действительно обусловить мышление хотя бы даже и в самой зачаточной степени.<sup>80</sup> Так что сторонникам подобных предположений остается, по всей видимости, уповать лишь на то, что всеми мыслительными ощущениями мы обязаны не чему иному, как банальной сложности<sup>81</sup> сопровождающих деятельность мозга вычислений (выполняющихся в соответствии с упомянутыми предположениями).

В связи с этим возникает еще несколько проблем, которых, насколько мне известно, всерьез пока не касался никто. Если предположить, что необходимым условием сознательной мыслительной деятельности является, главным образом, огромная сложность «соединений», формирующих в мозге сеть из взаимосвязанных нейронов и синапсов, то придется каким-то образом примириться и с тем, что сознание свойственно не всем отделам головного мозга человека в равной степени. Когда термин «мозг» употребляют без каких-либо уточнений, вполне естественно (по крайней мере, для неспециалиста) представлять себе обширные, покрытые извилинами внешние области, образующие так называемую кору головного мозга, – состоящий из серого вещества наружный слой головного мозга. В коре головного мозга содержится приблизительно сто тысяч миллионов ( $10^{11}$ ) нейронов, что и в самом деле дает ощутимый простор для формирования структур огромной сложности, однако кора – это еще далеко не весь мозг. В задней нижней части мозга находится еще один весьма важный сгусток спутанных нейронов, известный как мозжечок (см. рис. 1.6). Мозжечок, судя по всему, неким критическим образом связан с процессом выработки двигательных навыков; его действие можно наблюдать, когда человек овладевает тем или иным движением в совершенстве, т.е. когда движение перестает требовать сознательного обдумывания, как не требует обдумывания, скажем, ходьба. Сначала, когда мы еще только учимся какому-то новому навыку, нам необходимо контролировать свои действия сознательно, и этот контроль, по-видимому, требует существенного участия коры головного мозга. Однако впоследствии, по мере того, как необходимые движения становятся «автоматическими», управление ими постепенно переходит к мозжечку и осуществляется, по большей части, бессознательно. Учитывая, что деятельность мозжечка является, по всей видимости, абсолютно бессознательной, весьма примечателен тот факт, что количество нейронов в мозжечке может достигать половины того их количества, что содержится в коре головного мозга. Более того, именно в мозжечке располагаются такие нейроны, как клетки Пуркиньи (те самые, что имеют до 80'000 синаптических связей, о чем я уже упоминал в §1.2), так что общее

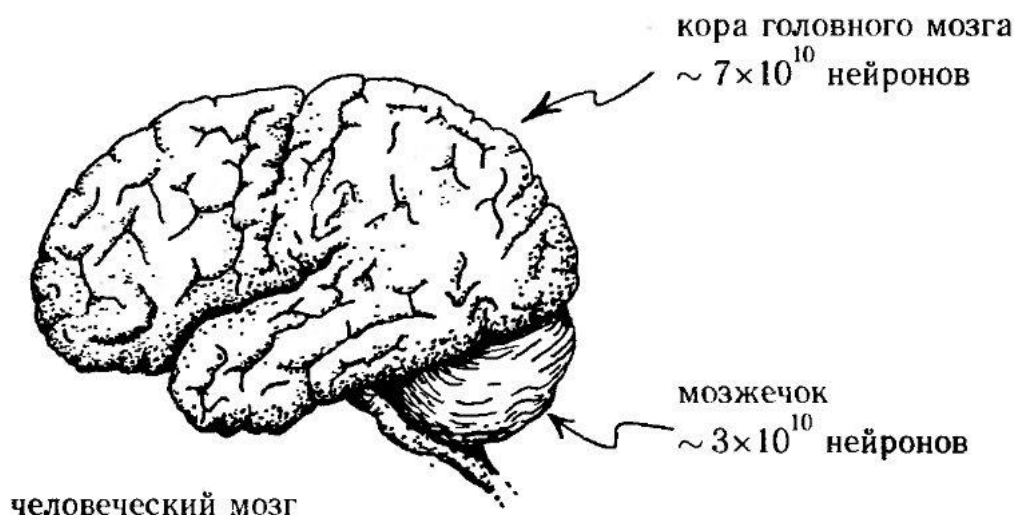
---

<sup>79</sup> В.Э.: Да – если мы встроим в компьютер способность улавливать электромагнитные волны, различать их по длине волны и особое внимание обращать на волны красного диапазона, и – далее – вложим этот аппарат в полную, интегральную систему, реализующую (а не имитирующую!) разум, то будет вполне ВСЁ, что требуется для «ощущения красного». И все объективные проявления в этом деле будут совпадать у этого искусственного интеллекта и у человека. Но субъективные... Они не наблюдаемы. Я, например, никогда не был уверен, что вижу «красное» таким же, как другие люди (и в юности меня это очень занимало, я даже писал об этом). Как это проверить? Как установить, что мое «красное» совпадает с тем «красным», что видит Пенроуз? Никак! Может быть, мы по-разному это видим, но только привыкли всю жизнь обозначать одним и тем же словом? И такая же ситуация будет и с тем компьютером ИИ. Иди-знай, что он там внутри чувствует – то же самое, или не то же самое?! А всё, что можно установить, измерить, проверить объективно – всё совпадает.

<sup>80</sup> В.Э.: Не знаю, как у представителей точки зрения *A*, но мне как стороннику точки зрения *C* представляется очевидным, что на современных компьютерах (правда, не промышленных для офисов, а специально сконструированных, но при теперешних технологиях) и при теперешних (только моих, а не любых) представлениях об ИИ, можно реализовать (не какой-то там зачаточный, а) полный интеллект. (Всё дело просто в том, что я знаю, КАК это надо делать, а Пенроуз не знает).

<sup>81</sup> В.Э.: Нет – отнюдь не на сложности (основные принципы интеллекта в общем-то довольно просты). Чтобы написать программу, извлекающую квадратный корень, нужно знать, КАК ее надо писать. Чтобы создать программу, реализующую разум, надо знать, КАК ее создавать. И в том, и в другом случае просто надо знать, КАКАЯ работа должна быть выполнена, КАК ее можно выполнить, что вообще из себя представляет тот объект, с которым мы работаем. Вот именно этого знания и не хватает у тех, кто систему, реализующую разум, написать не могут. (И это знание составляет сущность Веданской теории).

число связей между нейронами в мозжечке может оказаться ничуть не меньше аналогичного числа в головном мозге.<sup>82</sup> Если необходимым условием возникновения сознания считать одну лишь сложность нейронной сети, то неплохо было бы выяснить, почему же сознание никак, на первый взгляд, не проявляется в деятельности мозжечка. (Несколько дополнительных замечаний на эту тему приведены в §8.6).



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

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

### §1.15. Свидетельствуют ли ограниченные возможности сегодняшнего ИИ в пользу *C*?

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

<sup>82</sup> В.Э.: Для концептуального понимания разума (сознания и т.д.) вообще не нужны никакие сведения о физическом устройстве мозга (скорее, они мешают). Это так же, как для понимания программы извлечения квадратного корня вам не нужно знать, на каком компьютере она работает и как он устроен (и вообще она может работать на разных компьютерах разной конструкции). Точно так же системы, реализующие интеллект, могут работать на разных компьютерах (в мозге в том числе), и вовсе не важно физическое устройство этих компьютеров, а важны принципы работы, алгоритмы, структуры данных, информационные связи между программами и структурами.

научных терминах. Таким образом, для поддержания точки зрения  $\mathcal{C}$  необходимо найти примеры мыслительной деятельности, не поддающиеся никакому вычислению, и, кроме того, попытаться сообразить, как подобная деятельность может оказаться результатом тех или иных физических процессов. Остаток первой части моей книги будет направлен на достижение первой цели, во второй же части я представлю свои попытки продвинуться по направлению к цели номер два.

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

Как ни удивительно, главную неудачу современный искусственный интеллект терпит вовсе не в тех областях, где человеческий разум может вполне самостоятельно продемонстрировать поистине впечатляющую мощь – там, например, где отдельные люди-эксперты способны буквально потрясти всех окружающих какими-то своими специальными познаниями или способностью мгновенно выносить суждения, требующие крайне сложных вычислительных процедур, – а в вещах вполне «обыденных», какие на протяжении большей части своей сознательной жизни проделывают самые заурядные из представителей рода человеческого. Пока что ни один управляемый компьютером робот не может соперничать даже с малым ребенком в таком, например, простейшем деле, как сообразить, что для завершения рисунка необходим цветной карандаш, который валяется на полу в противоположном конце комнаты, после чего подойти к нему, взять и использовать по назначению.<sup>83</sup> Коли уж на то пошло, даже способности муравья, проявляющиеся в выполнении повседневной муравьиной работы, намного превосходят всё то, что можно реализовать с помощью самых сложных современных систем компьютерного управления. А с другой стороны, перед нами имеется поразительный пример способности компьютеров к чрезвычайно эффективным действиям – я имею в виду последние работы по созданию шахматных компьютеров. Шахматы, несомненно, представляют собой такой вид деятельности, в котором мощь человеческого интеллекта проявляется особенно ярко, хотя в полной мере эту мощь используют, к сожалению, лишь немногие. И всё же современные компьютерные системы играют в шахматы необычайно хорошо и способны выиграть у большинства шахматистов-людей. Даже лучшим из шахматистов приходится сейчас нелегко, и вряд ли им удастся надолго сохранить свое теперешнее превосходство над наиболее продвинутыми компьютерами.<sup>84</sup> Существует еще несколько узких областей, в которых компьютеры могут с успехом (постоянным или переменным) соперничать со специалистами-людьми. Кроме того, необходимо упомянуть и о таких видах интеллектуальной деятельности (например, о прямых численных расчетах), где способности компьютеров значительно превосходят способности людей.

---

<sup>83</sup> В.Э.: Не используют самопрограммирование. Даже муравей – самопрограммирующееся устройство, а в компьютерах этого сейчас никто не встраивает: – видимо, не умеют... Мысли вообще не идут в таком направлении. (Похоже, только у меня идут...)

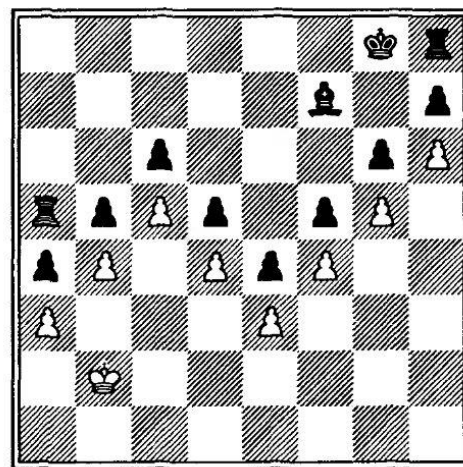
<sup>84</sup> См. [208].

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

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

Как мог столь искусный шахматист сделать такой очевидно глупый ход? Ответ заключается в следующем: помимо большого количества «позиций из учебника» программа *Deep Thought* содержала лишь инструкции, которые сводились исключительно к вычислению последовательности будущих ходов (на некоторую значительную глубину), позволяющей достичь максимального преимущества по фигурам. Ни на одном из этапов вычислений компьютер не обладал подлинным пониманием не только того, что может ему дать заслон из пешек, но и вообще любого из своих действий.<sup>85</sup>

Любой, кто в достаточной степени представляет себе общий принцип работы компьютера *Deep Thought* или других компьютерных систем для игры в шахматы, не станет удивляться тому, что эта система терпит крах в позициях вроде той, что показана на рис. 1.7. Мы не только способны понять в шахматах что-то такое, чего не понимает *Deep Thought*; мы, кроме того, кое-что понимаем и в процедурах (нисходящих), на которых построена вся работа *Deep Thought*, то есть мы способны как реально оценить, почему он сделал столь грубую ошибку, так и понять, почему в большинстве других случаев он может играть в шахматы настолько эффективно. Напрашивается, однако, вопрос: сможет ли *Deep Thought* или иная ИИ-система достичь когда-нибудь хоть какого-то подлинного понимания – подобного тому, каким обладаем мы сами – в шахматах или в чем-то еще? Некоторые сторонники ИИ скажут, что для обретения



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

<sup>85</sup> В.Э.: Проверку на ситуации типа «заслон из пешек» легко встроить в шахматную программу, но это, конечно, не сделает ее «понимающей». Чтобы сделать систему «понимающей», надо дать ей стартовый набор программ и потом путем самопрограммирования наращивать ее навыки. Но детально это невозможно разобрать здесь в сноске; читайте мои работы о системе Витос, которые сейчас существуют только на латышском, но готовятся к переводу на русский.

ИИ-системой «подлинного» понимания (что бы это ни значило) ее программа должна задействовать восходящие процедуры на гораздо более фундаментальном уровне, нежели это принято в программах теперешних шахматных компьютеров. Соответственно, в такой системе «понимание» развивалось бы постепенно по мере накопления «опыта», а не возникало бы в результате введения каких-то конкретных нисходящих алгоритмических правил. Нисходящие правила, достаточно простые и прозрачные, не способны сами по себе обеспечить вычислительную основу для подлинного понимания, поскольку само понимание этих правил позволяет нам осознать их фундаментальные ограничения.

Этот момент мы более подробно рассмотрим в главах 2 и 3. А что же в самом деле восходящие вычислительные процедуры? Могут ли они составить основу для понимания? В главе 3 я приведу рассуждения, доказывающие обратное. Пока же мы можем просто взять на заметку тот факт, что современные компьютерные системы восходящего типа никоим образом не обеспечивают замены подлинному человеческому пониманию ни в одной из важных областей интеллектуальной компетенции, требующих настоящего живого человеческого понимания и интуиции. Такую позицию, я уверен, сегодня разделяют многие. Весьма оптимистичные перспективы,<sup>86</sup> время от времени выдвигаемые сторонниками идеи искусственного интеллекта и производителями экспертных систем, пока что в большинстве своем реализованы не были.

Однако в том, что касается возможных результатов развития искусственного интеллекта, мы всё еще находимся в самом начале пути. Сторонники ИИ (в форме *A* или *B*) уверяют нас, что проявление существенных элементов понимания в поведении их систем с компьютерным управлением – всего лишь вопрос времени и, быть может, некоторых, пусть и значительных, технических усовершенствований. Несколько позднее я попробую поспорить с этим заявлением в более точных терминах, опираясь на то, что некие фундаментальные ограничения присущи любой чисто вычислительной системе, будь она нисходящей или восходящей. Не исключая возможности того, что, будучи достаточно грамотно сконструированной, такая система сможет в течение некоторого продолжительного периода времени поддерживать иллюзию обладания чем-то, подобным пониманию (как это произошло с компьютером *Deep Thought*), я всё же утверждаю, что на деле полная ее неспособность к пониманию в общем смысле этого слова непременно в конце концов обнаружится – по крайней мере, в принципе.<sup>87</sup>

Для приведения точных аргументов мне придется обратиться к математике, причем я намерен показать, что к одним лишь вычислениям невозможно свести даже математическое понимание.<sup>88</sup> Некоторые защитники ИИ могут счесть это весьма удивительным, ибо они утверждают,<sup>89</sup> что те способности, которые сформировались в процессе эволюционного развития человека сравнительно недавно (например, способность выполнять арифметические или алгебраические вычисления), «осваиваются» компьютерами легче всего, и именно в этих областях компьютеры на настоящий момент значительно опережают «человека вычисляющего»; овладение же теми способностями, что развились в начале эволюционного пути – такими, например, как умение ходить или интерпретировать сложные визуальные сцены, – не требует практически никакого труда от человека, тогда как сегодняшние компьютеры даже при всем старании демонстрируют в этом «виде спорта» весьма посредственные результаты. Я рассуждаю несколько иначе. Современный компьютер легко справится с любой сложной деятельностью – будь то математические вычисления, игра в шахматы или выполнение какой-либо работы по дому, – но лишь при условии, что эту деятельность можно описать в виде набора четких вычислительных правил; а вот собственно понимание, лежащее в основе этих самых вычислительных правил, оказывается феноменом, для вычисления недоступным.

---

<sup>86</sup> См. [124].

<sup>87</sup> В.Э.: Конечно, обнаружится, если это имитация интеллекта, то есть если система работает НЕ по тем принципам, по которым работает интеллект. Конечно, «полная неспособность к пониманию» НЕ обнаружится, если система будет создана действительно понимающей, то есть будет реализован настоящий интеллект по тем принципам, по которым интеллект вообще работает.

<sup>88</sup> В.Э.: А я намерен показать, что можно свести – только не теми способами, о которых думает и рассуждает Пенроуз, а другими, о которых он и представления не имеет.

<sup>89</sup> См., например, [268].

### §1.16. Доказательство на основании теоремы Гёделя

Как можем мы быть уверены в том, что вышеописанное понимание не может, в сущности, быть сведено к набору вычислительных правил? Несколько позже (в главах 2 и 3) я приведу некоторые очень серьезные доводы в пользу того, что проявления понимания (по крайней мере, определенных его видов) невозможно достоверно моделировать посредством каких угодно вычислений – ни нисходящего, ни восходящего типа, ни любой из их комбинаций. Таким образом, за реализацию присущей человеку способности к «пониманию» должна отвечать какая-то невычислительная деятельность мозга или разума. Напомним, что термином «невычислительный» в данном контексте (см. §1.5, §1.9) мы характеризуем феномен, который невозможно эффективно моделировать с помощью какого угодно компьютера, основанного на логических принципах, общих для всех современных электронных или механических вычислительных устройств. При этом термин «невычислительная активность» вовсе не предполагает невозможности описать такую активность научными и, в частности, математическими методами. Он предполагает лишь то, что точки зрения  $\mathcal{A}$  и  $\mathcal{B}$  оказываются не в состоянии объяснить, каким именно образом мы выполняем все те действия, которые представляют собой результат сознательной мыслительной деятельности.<sup>90</sup>

Существует, по меньшей мере, логическая возможность того, что обладающий сознанием мозг (или сознательный разум) может функционировать в соответствии с такими невычислительными законами (см. §1.9). Однако так ли это? Представленные в следующей главе (§2.5) рассуждения содержат, как мне кажется, весьма четкое доказательство наличия в нашем сознательном мышлении невычислительной составляющей. Основаны эти рассуждения на знаменитой и мощной теореме математической логики, сформулированной великим логиком, чехом<sup>91</sup> по происхождению, Куртом Гёделем. Для моих целей будет вполне достаточно существенно упрощенного варианта этой теоремы, который не потребует от читателя слишком обширных познаний в математике (что касается математики, то я также позаимствую кое-что из одной важной идеи, высказанной несколько позднее Аланом Тьюрингом). Любой достаточно серьезно настроенный читатель без труда разберется в моих рассуждениях. Доказательства гёделевского типа, да еще и примененные в подобном контексте, подвергаются время от времени решительным нападкам.<sup>92</sup> Вследствие этого у некоторых читателей может сложиться впечатление, что мое основанное на теореме Гёделя доказательство было полностью опровергнуто. Должен заметить, что это далеко не так. За прошедшие годы действительно выдвигалось множество контраргументов. Мишенью для многих из них послужило одно из самых первых таких доказательств (направленное в поддержку ментализма и против физикализма), предложенное оксфордским философом Джоном Лукасом [246]. Опираясь на результаты теоремы Гёделя, Лукас доказывал, что мыслительные процессы невозможно воспроизвести вычислительными методами. (Подобные соображения выдвигались и ранее; см., например, [271].) Мое доказательство, пусть и построенное на том же фундаменте, выдержано всё же в несколько ином духе, нежели доказательство Лукаса; кроме того, в число моих задач не входила неременная поддержка ментализма. Я думаю, что моя формулировка способна лучше противостоять различным критическим замечаниям, выдвинутому в свое время против доказательства Лукаса, и во многих отношениях выявить их несостоятельность. Ниже (в главах 2 и 3) мы подробно рассмотрим все контраргументы, которые когда-либо попадались мне на глаза. Надеюсь, что мои сопутствующие комментарии не только помогут прояснить некоторые, похоже, широко распространенные заблуждения относительно смысла доказательства Гёделя, но и дополнят, по-видимому, неудовлетворительно краткое рассмотрение этого вопроса, предпринятое в НРК. Я намерен показать, что большая часть этих контраргументов произрастает,

<sup>90</sup> В.Э.: Это объясняет точка зрения  $\mathcal{C}$ .

<sup>91</sup> В.Э.: Уж кем-кем, а чехом он не был. (По национальности он был немцем – не знаю, может с примесью еще какой-нибудь крови; если да, то, скорее всего, еврейской, потому что всю жизнь дружил с евреями; – родился он в Австро-Венгрии на территории будущей Чехословакии, но когда такое государство действительно образовалось, то он уехал из него в Австрию и в 23 года принял австрийское гражданство).

<sup>92</sup> О доказательстве Лукаса см. [320], [345], [25], [163], [164], [236], [237], [202], [38]; см. также [247]. Что касается моей версии, кратко представленной в НРК, с. 416–418, то где только ее не критиковали: см., в особенности, [344] и многочисленные статьи в *Behavioral and Brain Sciences*: [37], [43], [47], [73], [74], [80], [97], [154], [199], [220], [251], [250], [253], [269], [307], [324], [366], [386]; мои ответы на критику см. в [292], [298] и [178]; см. также [95], [294].

в сущности, из банальных недоразумений, тогда как остальные, основанные на более или менее осмысленных и требующих детального рассмотрения возражениях, представляют собой, в лучшем случае, не более чем возможные «лазейки» в духе взглядов  $\mathcal{A}$  или  $\mathcal{B}$ ; при этом они не дают – в чем у нас еще будет возможность убедиться – сколько-нибудь правдоподобного объяснения действительным последствиям наличия у нас способности «понимать», да и в любом случае эти лазейки не представляют особой ценности для развития идеи ИИ. Так что тем, кто по-прежнему полагает, что все внешние проявления процессов сознательного мышления можно адекватно воспроизвести вычислительными методами, в рамках положений  $\mathcal{A}$  или  $\mathcal{B}$ , я могу лишь порекомендовать повнимательнее следить за предлагаемой ниже аргументацией.<sup>93</sup>

### §1.17. Платонизм или мистицизм?

Критики, впрочем, могут возразить, что отдельные выводы в рамках этого доказательства Гёделя следует рассматривать не иначе как «мистические», поскольку упомянутое доказательство, судя по всему, вынуждает нас принять либо точку зрения  $\mathcal{C}$ , либо точку зрения  $\mathcal{D}$ ; подобный взгляд, разумеется, не более приемлем, нежели любая из вышеупомянутых лазеек, полученных из теоремы Гёделя. Что касается  $\mathcal{D}$ , то здесь я, вообще говоря, полностью с критиками согласен. Мои собственные причины неприятия  $\mathcal{D}$  – точки зрения, настаивающей на полном бессилии науки перед тайною разума, – проистекают из осознания того факта, что только благодаря применению научных и, в частности, математических методов был достигнут хоть какой-то реальный прогресс в понимании происходящих в окружающем нас мире процессов. Более того, если мы и располагаем какими-то достоверными сведениями о разуме, то только о том разуме, который тесно связан с конкретным физическим объектом – мозгом, – причем различным состояниям разума четко соответствуют различные физические состояния мозга. По всей видимости, с теми или иными специфическими типами физической активности мозга можно ассоциировать и психические состояния сознания. Если бы не таинственные аспекты сознания, связанные с формированием «осознания» и, быть может, с проявлениями «свободы воли», которые пока что не поддаются физическому описанию, нам бы и в голову не пришло, что для объяснения разума, являющегося по всем признакам продуктом протекающих внутри мозга физических процессов, стандартных научных методов может и не хватить.<sup>94</sup>

С другой стороны, следует понимать, что наука (и, в частности, математика) и сама по себе являет нам мир, исполненный тайн. Чем глубже мы проникаем в процессе научного познания в суть вещей, тем более фундаментальные тайны открываются нашему взору. Быть может, стоит в этой связи упомянуть и о том, что физики, более непосредственно знакомые с головоломной и непостижимой манерой, в какой реально проявляет себя материя, склонны видеть мир в менее классически механистическом свете, нежели биологи. В главе 5 мы поговорим о некоторых наиболее таинственных аспектах квантового поведения, обнаруженных относительно недавно. Возможно, для полного «охвата» тайны разума нам придется несколько расширить границы того, что мы в настоящее время называем наукой, однако я не вижу причин напрочь отказываться от тех методов, которые так замечательно служили нам до сих пор. Таким образом, если гёделевские соображения подталкивают нас к принятию точки зрения  $\mathcal{C}$  в том или ином ее виде (а я полагаю, что так оно и есть), то нам поневоле придется принять и некоторые другие ее следствия. Иными словами, следуя этим путем, мы приходим, ни много ни мало, к объективному идеализму по Платону. Согласно учению Платона, математические концепции и математические истины существуют в их собственном, вполне реальном мире, в котором отсутствует течение времени и который не имеет физического местонахождения. Мир Платона – это идеальный мир совершенных форм, отличный от физического мира, но являющийся основой для его понимания. Он, кроме того, никак не связан с нашими несовершенными мысленными построениями, однако человеческий разум способен получить в некотором смысле непосредственный доступ в это платоново царство благодаря способности «осознавать» математические формы и рассуждать о них. Нашему «платоническому» восприятию, как вскоре выяснится, может иногда поспособствовать вычисление, однако в общем это восприятие вычислением не ограничено. Согласно такому платоническому подходу, именно способность «осознавать» математические концепции дает

<sup>93</sup> В.Э.: Разберем, разберем, за нами не заржавеет...

<sup>94</sup> В.Э.: Но хватает.

разуму мощь, далеко превосходящую всё, чего можно добиться от устройства, работа которого основывается исключительно на вычислении.<sup>95</sup>

### §1.18. Почему именно математическое понимание?

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

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

Почему же в таком случае я решил рассмотреть вопрос сознания прежде всего в математическом контексте? Причина заключается в том, что только в математических рамках мы можем рассчитывать на возможность хоть сколько-нибудь строгой демонстрации непременной

---

<sup>95</sup> В.Э.: Здесь Пенроуз прав, но только он сам не понимает, в КАКОМ смысле он прав, и думает, что он прав в совсем другом смысле. Так называемый «мир идей Платона» – это на самом деле «мир» потенциальных продуктов мозговых программ (в общем случае: вообще программ интеллекта); и действительно именно «способность осознавать» (т.е. работать с объектами этого «мира») «дает разуму мощь, далеко превосходящую всё, чего можно добиться от устройства, работа которого основывается исключительно на вычислении», как сказал Пенроуз, а мы скажем точнее: «дает устройству, работа которого основывается исключительно на вычислении, ту мощь, которая называется разумом». Да! – именно в «мире Платона» ключ к разгадке интеллекта: кто разгадал «идеи» Платона, тот разгадал уже и сам «разум».

невыхислимости, по крайней мере, некоторой части сознательной деятельности. Вопрос вычислимости по самой своей природе является, безусловно, математическим. Нельзя ожидать, что нам удастся дать хоть какое-то «доказательство» невычислимости того или иного процесса, не обратившись при этом к математике. Я хочу убедить читателя в том, что всё, что мы делаем нашим мозгом или разумом в процессе понимания математического суждения, существенно отличается от того, чего мы можем добиться от какого угодно компьютера; если мне это удастся, то читателю будет намного легче оценить роль невычислительных процессов в сознательном мышлении вообще.

А разве не очевидно, возразят мне, что восприятие того же красного цвета никак не может быть вызвано просто выполнением какого бы то ни было вычисления. К чему вообще утруждать себя какими-то ненужными математическими демонстрациями, когда и без того совершенно ясно, что *qualia* – т.е. субъективные ощущения – никак не связаны с вычислениями<sup>96</sup>? Один из ответов заключается в том, что такое доказательство от «очевидного» (как бы благожелательно я ни относился к подобному способу доказательства) применимо только к пассивным аспектам сознания. Как и китайскую комнату Серла, его можно представить в качестве аргумента против точки зрения  $\mathcal{A}$ , а вот между  $\mathcal{C}$  и  $\mathcal{B}$  разницы для него не существует.

Более того, мне представляется крайне уместным побить функционалистов вместе с их вычислительной моделью (т.е. точкой зрения  $\mathcal{A}$ ), так сказать, на их собственном поле; ведь это именно функционалисты настаивают на том, что все *qualia* на самом деле должны быть так или иначе обусловлены банальным выполнением соответствующих вычислений, невзирая на то, сколь невероятной такая картина может показаться на первый взгляд. Ибо, аргументируют они, что же еще можем мы эффективно делать своим мозгом, как не выполнять те или иные вычисления? Для чего вообще нужен мозг, если не в качестве своеобразной системы управления вычислениями – да, чрезвычайно сложными, но всё же вычислениями? Какие бы «ощущения осознания» ни пробуждались в нас в результате той или иной функциональной активности мозга, эти ощущения, согласно функционалистской модели, непременно являются результатом некоторой вычислительной процедуры. Функционалисты любят упрекать тех, кто не признает за вычислительной моделью способности объяснить любые проявления активности мозга, включая и сознание, в склонности к мистицизму. (Надо понимать так, что единственной альтернативой точки зрения  $\mathcal{A}$  является  $\mathcal{D}$ .)

Во второй части книги я намерен привести несколько частных предположений относительно того, что еще может вполне эффективно делать мозг, допускающий научное описание. Не стану отрицать, некоторые «конструктивные» моменты моего доказательства являются чисто умозрительными. И всё же я полагаю, что мои доводы в пользу невычислимости хотя бы некоторых мыслительных процессов весьма убедительны; а для того, чтобы эта убедительность переросла в неотразимость, их следует применить к математическому мышлению.

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

Допустим однако, что мы все уже согласны с тем, что при формировании осознанных математических суждений и получении осознанных же математических решений в нашем мозге действительно происходит что-то невычислимое. Каким образом это поможет нам понять причины ограниченных способностей роботов, которые, как я упоминал ранее, значительно хуже справляются с элементарными, «бытовыми», действиями, нежели со сложными задачами, для выполнения которых требуются высококвалифицированные специалисты-люди? На первый взгляд, создается впечатление, что мои выводы в корне противоположны тем, к которым придет всякий здравомыслящий человек, исходя из известных ограничений искусственного интеллекта – по крайней мере, сегодняшних ограничений. Ибо многим почему-то кажется, что я утверждаю, будто невычислимое поведение должно быть связано скорее с пониманием крайне сложных областей математики, а никак не с обыденным, бытовым поведением. Это не так. Я утверждаю лишь, что пониманию сопутствуют невычислимые процессы одинаковой природы, вне зависимости от того, идет ли речь о подлинно математическом восприятии, скажем, бесконеч-

---

<sup>96</sup> В.Э.: Совершенно ясно, что «ощущение красного» связано с «чисто вычислительной» способностью компьютера 1) обнаруживать электромагнитные волны, 2) различать их по длине волны и 3) особо выделять один диапазон (называемый красным цветом).

ного множества натуральных чисел или всего лишь об осознании того факта, что предметом удлинённой формы можно подпереть открытое окно, о понимании того, какие именно манипуляции следует произвести с куском веревки для того, чтобы привязать или, напротив, отвязать уже привязанное животное, о постижении смысла слов «счастье», «битва» или «завтра» и, наконец, о логическом умозаключении относительно вероятного местонахождения правой ноги Авраама Линкольна, если известно, что левая его нога пребывает в настоящий момент в Вашингтоне, – я привел здесь некоторые из примеров, оказавшихся на удивление мучительными для одной реально существующей ИИ-системы!<sup>97</sup> Такого рода невычислимые процессы лежат в основе всякой деятельности, результатом которой является непосредственное осознание чего-либо. Именно это осознание позволяет нам визуализировать геометрию движения деревянного бруска, топологические свойства куска веревки или же «связность» Авраама Линкольна. Оно также позволяет нам получить до некоторой степени прямой доступ к опыту другого человека, с помощью чего мы можем «узнать», что этот другой, скорее всего, подразумевает под такими словами, как «счастье», «битва» и «завтра», несмотря даже на то, что предлагаемые в процессе общения объяснения зачастую оказываются недостаточно адекватными. Передать «смысл» слов от человека к человеку всё же возможно, однако не с помощью объяснений различной степени адекватности, а лишь благодаря тому, что собеседник уже, как правило, имеет в сознании некий общий образ возможного смысла этих слов (т.е. «осознает» их), так что даже очень неадекватных объяснений обычно бывает вполне достаточно для того, чтобы человек смог «уловить» верный смысл. Именно наличие такого общего «осознания» делает возможным общение между людьми. И именно этот факт ставит неразумного, управляемого компьютером робота в крайне невыгодное положение. (В самом деле, уже самый смысл понятия «смысл слова» изначально воспринимается нами как нечто само собой разумеющееся, и поэтому совершенно непонятно, каким образом такое понятие можно сколько-нибудь адекватно описать нашему неразумному роботу.) Смысл можно передать лишь от человека к человеку, потому что все люди имеют схожий жизненный опыт или аналогичное внутреннее ощущение «природы вещей». Можно представить «жизненный опыт» в виде своеобразного хранилища, в которое складывается память обо всем, что происходит с человеком в течение жизни, и предположить, что нашего робота не так уж и сложно таким хранилищем оснастить. Однако я утверждаю, что это не так; ключевым моментом здесь является то, что рассматриваемый субъект, будь то человек или робот, должен свой жизненный опыт осознавать.

Что же заставляет меня утверждать, будто упомянутое осознание, что бы оно из себя ни представляло, должно быть невычислимым – иначе говоря, таким, что его не сможет ни достичь, ни хотя бы воспроизвести ни один робот, управляемый компьютером, построенным исключительно на базе стандартных логических концепций машины Тьюринга (или эквивалентной ей) нисходящего либо восходящего типа? Именно здесь и играют решающую роль гёделевские соображения. Вряд ли мы в настоящее время можем многое сказать об «осознании», например, красного цвета; а вот относительно осознания бесконечности множества натуральных чисел кое-что определенное нам таки известно. Это такое «осознание», благодаря которому ребенок «знает», что означают слова «ноль», «один», «два», «три», «четыре» и т.д. и что следует понимать под бесконечностью этой последовательности, хотя объяснения ему были даны до нелепости ограниченные и, на первый взгляд, к делу почти не относящиеся, на примере нескольких бананов и апельсинов. Из таких частных примеров ребенок и в самом деле способен вывести абстрактное понятие числа «три». Более того, он также оказывается в состоянии понять, что это понятие является лишь звеном в бесконечной цепочке похожих понятий («четыре», «пять», «шесть» и т.д.). В некотором платоническом смысле ребенок изначально «знает», что такое натуральные числа.<sup>98</sup>

<sup>97</sup> Примеры взяты из какой-то английской телевизионной программы; возможно, из «Машины мечты» (*The Dream Machine*, декабрь 1991 г.) – четвертой из цикла программ ВВС «Мыслящая машина» (*The Thinking Machine*). О последних достижениях в области «искусственного понимания», а в особенности, о захватывающем проекте Дугласа Лената «СУС» можно прочесть в [124].

<sup>98</sup> В.Э.: Ребенок «изначально знает», «что такое натуральные числа» тогда, когда в мозге у него создана программа классификации множеств внешнего мира по количеству элементов (назовем ее программой N). Если эта мозговая программа у него есть (т.е. он умеет классифицировать множества), то он (стандартным для программ интеллекта способом) строит «номиналии» (так эти внутрикомпьютерные объекты называются в Веданской теории) потенциальных продуктов программы N (таксонов классификации) и может ими в дальнейшем оперировать (это и есть натуральные числа). Немножко

Возможно, кто-то усмотрит здесь некий налет мистики, однако в действительности мистика здесь не при чем. Для понимания последующих рассуждений крайне важно отличать такое платоническое знание от мистицизма. Понятия, «известные» нам в платоническом смысле, суть вещи для нас «очевидные»: вещи, которые сводятся к воспринятому когда-то «здоровому смыслу», – при этом мы не можем охарактеризовать эти понятия во всей их полноте посредством вычислительных правил. Действительно – и это станет ясно из дальнейших рассуждений, связанных с доказательством Гёделя, – не существует способа целиком и полностью охарактеризовать свойства натуральных чисел на основе лишь таких правил.<sup>99</sup> А как же тогда описания числа через яблоки или бананы дают ребенку понять, что означают слова «три дня», и откуда ему знать, что смысл абстрактного понятия числа «три» здесь совершенно тот же, что и в словах «три апельсина»<sup>100</sup>? Разумеется, такое понимание иногда приходит к ребенку далеко не сразу, и на первых порах он, бывает, ошибается, однако суть не в этом. Суть в том, что подобное осознание вообще возможно. Абстрактное понятие числа «три», равно как и представление о том, что существует бесконечная последовательность аналогичных понятий – собственно последовательность натуральных чисел, – и в самом деле вполне доступно человеческому пониманию, однако, повторяю, лишь через осознание.

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

Важный момент, как мне кажется, заключается в том, что визуализация содержит некий элемент оценки того, что человек видит, то есть сопровождается пониманием. Чтобы проиллюстрировать, что я имею в виду, давайте рассмотрим одно элементарное арифметическое правило,

---

поисследовав свою программу N, он находит, что никакой таксон для нее не будет последним (т.е. что натуральных чисел бесконечно много). Так он «осознал» и бесконечность – и всё это делает компьютер (у ребенка мозговой, но может и другой). Сущность в том, что имеется программа N, а другие программы ее анализируют со стороны – осуществляя «понимание», «осознание» и т.д. ее самой, ее продуктов и даже бесконечности ее продуктов. Эта работа одной программы с другой программой – сущность разума, тот самый недостающий ингредиент, который Пенроуз не находит в рамках классической науки и поэтому хочет искать в квантовой механике.

<sup>99</sup> В.Э.: Я только что охарактеризовал; правда, здесь, в сноске, очень бегло, но у меня есть и целые книги об этом. Я показал, КАК компьютеры рожают понятия чисел (не только натуральных, но и вообще всех), и КАК они могут оперировать с абстрактными числами (и вообще объектами). Здесь опять видна еще одна особенность пенроузовского (и вообще математического) восприятия мира. Я указал мозговую программу (N), которая порождает у людей «понятие» о натуральных числах. Исследования чисел есть на самом деле исследования этой программы (ее потенциальных продуктов). Разумеется, сама эта программа не предоставляет готовые сведения о том, каковы свойства чисел, – их надо исследовать. Пенроуз же не подозревает о существовании программы N. Поэтому для него числа – объекты Платонова мира, а свойства их определяются аксиомами и не могут быть определены полностью – как это следует из теоремы Гёделя. Но не аксиомами определяются свойства чисел – а программой N! И поэтому теорема Гёделя тут не при чем. Пенроуз думает об объектах, которые определяют его аксиомы: «мы не можем охарактеризовать эти понятия во всей их полноте посредством вычислительных правил». А на самом деле эти «понятия» есть именно продукты мозговой ПРОГРАММЫ!

<sup>100</sup> В.Э.: Очень просто – его программа классификации N относит и объект «три дня», и объект «три апельсина» к одному и тому же таксону классификации множеств.

а именно: для любых двух натуральных чисел (т.е. неотрицательных целых чисел 0, 1, 2, 3, 4,...)  $a$  и  $b$  справедливо следующее равенство:

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

Следует пояснить, что это высказывание не является пустым, хотя части уравнения и имеют различный смысл.<sup>101</sup> Запись  $a \times b$  слева означает совокупность  $a$  групп по  $b$  объектов в каждой;  $b \times a$  справа –  $b$  групп по  $a$  объектов в каждой. В частном случае, например, при  $a = 3$  и  $b = 5$  запись  $a \times b$  можно представить следующим рядом точек:

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

в то время как для  $b \times a$  имеем

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

Общее число точек в каждом случае одинаково, следовательно, справедливо равенство

$$3 \times 5 = 5 \times 3.$$

В истинности этого равенства можно удостовериться, представив зрительно матрицу

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

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

Важный момент описанной визуализации заключается в том, что она непосредственно дает нам нечто гораздо более общее, чем просто частное численное равенство  $3 \times 5 = 5 \times 3$ . Иными словами, в конкретных числовых значениях  $a = 3$  и  $b = 5$ , участвующих в данной процедуре, нет ничего особенного. Полученное правило будет применимо, даже если, скажем,  $a = 79'797'000'222$ , а  $b = 50'000'123'555$ , и мы с уверенностью можем утверждать, что  $79'797'000'222 \times 50'000'123'555 = 50'000'123'555 \times 79'797'000'222$ , несмотря на то, что у нас нет ни малейшей возможности сколько-нибудь точно представить себе визуально прямоугольную матрицу такого размера (да и ни один современный компьютер не сможет перечислить все ее элементы). Мы вполне можем заключить, что вышеприведенное равенство должно быть истинным – или что истинным должно быть равенство общего вида<sup>102</sup>  $a \times b = b \times a$  – на основании, в сущности, той же самой визуализации,<sup>103</sup> которую мы применяли для конкретного случая  $3 \times 5 = 5 \times 3$ . Нужно просто несколько «размыслить» мысленно действительное количество строк и столбцов рассматриваемой матрицы, и равенство становится очевидным.

<sup>101</sup> В.Э.: Видимо, неправильный перевод; должно быть: «... потому что части уравнения имеют различный смысл».

<sup>102</sup> Необходимо отметить, что это равенство не является истинным для различных странных «чисел», встречающихся порой в математике, например, для трансфинитных чисел, о которых упоминается в пояснении к Q19, §2.10. Однако для натуральных чисел, о которых здесь, собственно, и идет речь, оно всегда справедливо.

<sup>103</sup> В.Э.: Ну, здесь вообще-то дело обстоит не так. Пенроуз так акцентирует «визуализацию», чтобы подойти к «очевидному» (латинский и далее английский корень у обоих слов один), но это он делает потому, что не подозревает о существовании программы N, которая задействована здесь самым прямым образом. Сама по себе «визуализация» (и то, что на этой странице мы действительно видим матрицу) еще ничего не дает (кошки и собаки тоже ее видят, но к математике не приходят). Важно, что эту картину в мозге начинают обрабатывать программой N, и тогда убеждаются, что если пускать N сначала по столбцам или сначала по строкам, то при обоих способах получится один и тот же результат. Формула  $3 \times 5 = 5 \times 3$  кодирует не что иное, как эквивалентность двух различных способов запуска программы N (двух групп измерений, сделанных ею). А формула  $a \times b = b \times a$  является не результатом «размытой визуализации», а кодировкой того (найденного мозгом при исследовании программы N) факта, что эти два способа измерения останутся эквивалентными при всех измеряемых конкретных множествах, независимо от количества элементов в них. Так что здесь мы имеем дело с изучением мозговой программы N.

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

Кое-кто, наверное, уже вообразил, что в таком случае можно раз и навсегда составить список всех «возможных» этапов умозаключений и тогда всякое доказательство можно будет свести к вычислению, т.е. к простым механическим манипуляциям полученными очевидными этапами. Доказательство Гёделя (§2.5) как раз и демонстрирует невозможность реализации такой процедуры. Нельзя совершенно избавиться от необходимости в новых «очевидно понимаемых» отношениях. Таким образом, математическое понимание никоим образом не сводится к бездумному вычислению.<sup>104</sup>

### §1.20. Мысленная визуализация и виртуальная реальность

Интуитивные математические процедуры,<sup>105</sup> описанные в §1.19, имеют весьма ярко выраженный специфический геометрический характер.<sup>106</sup> В математических доказательствах применяются и многие другие типы интуитивных процедур, причем некоторые из них весьма далеки от «геометричности». Однако, как показывает практика, геометрические интуитивные представления чаще всего дают более глубокое математическое понимание. Полагаю, было бы весьма полезно выяснить, какие же именно физические процессы происходят в нашем мозге, когда мы визуализируем что-либо геометрически. Начнем хотя бы с того, что никакой логической необходимости в том, чтобы непосредственным результатом этих процессов было «геометрическое отражение» визуализируемого объекта, по сути дела, не существует. Как мы увидим далее, здесь может получиться нечто совсем иное.

Здесь уместно провести аналогию с феноменом, именуемым «виртуальной реальностью». Феномен этот, согласно распространенному мнению, имеет самое прямое отношение к теме «визуализации». Методы виртуальной реальности<sup>107</sup> позволяют создать компьютерную модель какой-либо не существующей в природе структуры, – например, здания на стадии архитектурного проекта, – затем модель проецируется в глаз наблюдателя-человека, который, предположительно, воспринимает ее как «реальное» здание. Совершая движения глазами, головой или, может быть, ногами, словно прогуливаясь вокруг демонстрируемого ему здания, наблюдатель может разглядывать его с разных сторон – точно так же, как если бы здание действительно было реальным (см. рис. 1.8). Согласно некоторым предположениям,<sup>108</sup> выполняемые мозгом в процессе сознательной визуализации операции (какой бы ни была их истинная природа) аналогичны вычислениям, производимым при построении такой виртуальной модели. В самом деле, мысленно осматривая какую-то реально существующую неподвижную структуру, человек, по всей видимости, создает в уме некую модель, которая остается неизменной, несмотря на постоянные движения его головы, глаз и тела, приводящие к непрерывной смене образов, возникающих на сетчатке его глаз. Такие поправки на движения тела играют весьма существенную роль при построении виртуальной реальности, и высказывались предположения в

<sup>104</sup> В.Э.: «Математическое понимание» сводится не к «бездумному вычислению», а к изучению мозговых программ (не всех, а определенной группы, составляющей истинный предмет математики), как это я показал выше в данном самим Пенроузом примере и на программе N.

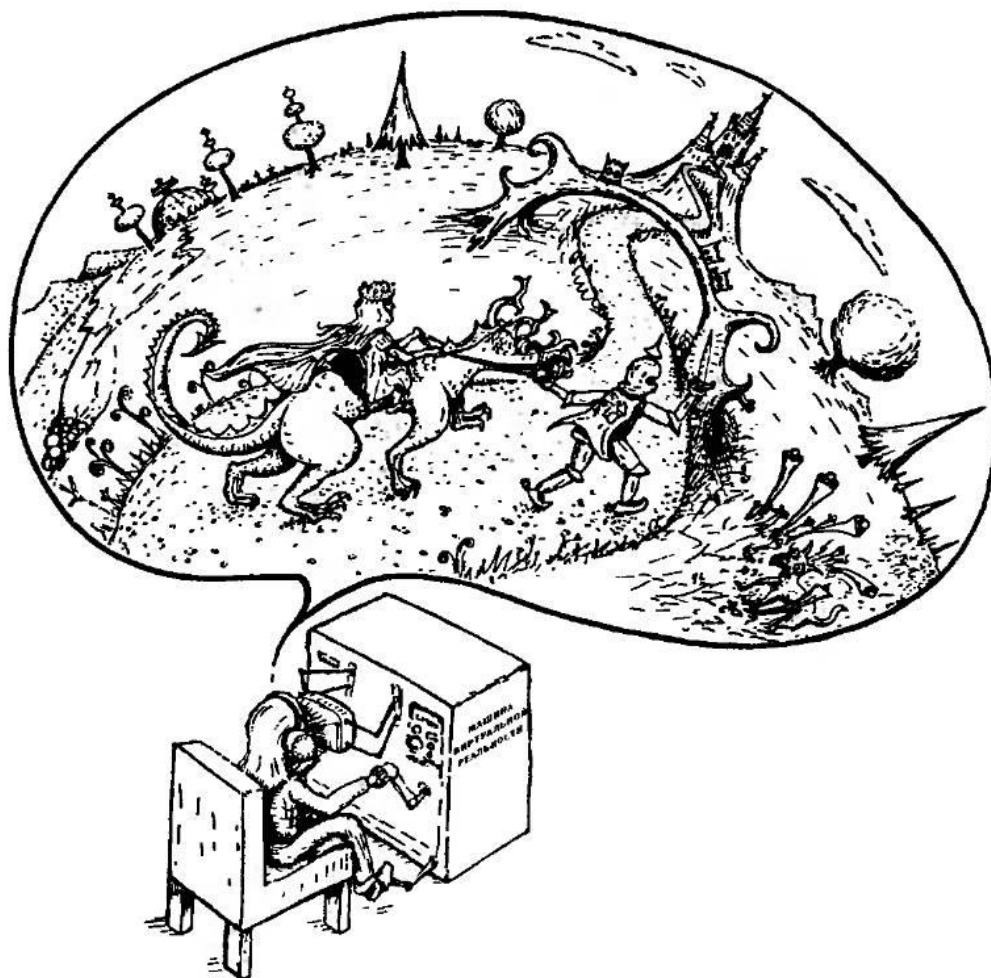
<sup>105</sup> В.Э.: Термин «интуитивные математические процедуры» неловок и свидетельствует о непонимании говорящим действительной природы вещей; лучше говорить об изучаемых математикой мозговых программах (таких как N) – тогда всё станет на свои места и не надо будет блуждать в потемках.

<sup>106</sup> В.Э.: Ну, рассмотренные в §1.19 примеры имели «геометрический характер» лишь постольку, поскольку измеряемые множества «визуализировались» перед тем, как их преподнести программе N. Сама программа N не имеет «геометрической природы».

<sup>107</sup> Весьма живо и популярно всё это описано в [389].

<sup>108</sup> Подобное предположение выдвинул, например, Ричард Доукинс в своих «Рождественских лекциях» (BBC, 1992 г.).

том смысле, что нечто подобное должно происходить и при создании «мысленных моделей», представляющих собой результаты актов визуализации. Такие вычисления, разумеется, вовсе не обязаны иметь целью воспроизведение реальной геометрической структуры моделируемой конструкции (или ее «отражение»). Сторонникам точки зрения *A* в таком случае пришлось бы рассматривать сознательную визуализацию как результат своего рода численного моделирования окружающего мира в голове человека. Я же полагаю, что всякий раз, когда мы сознательно воспринимаем ту или иную визуальную сцену, сопровождающее этот процесс понимание представляет собой нечто, существенно отличное от моделирования мира методами вычислительного характера.



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

Можно также предположить, что внутри мозга функционирует нечто вроде «аналогового компьютера», в котором моделирование внешнего мира реализуется не с помощью цифровых вычислений, как в современных электронных компьютерах, а с помощью некоторой внутренней структуры, физическое поведение которой каким-то однозначным образом отражает поведение моделируемой внешней системы. Допустим, например, что нам необходимо аналоговое устройство<sup>109</sup> для моделирования движений некоторого внешнего твердого тела. Для создания такого устройства мы, очевидно, воспользуемся весьма простым и естественным способом. Мы отыщем внутри системы реальное физическое тело той же формы (но меньшего размера), что и моделируемый внешний объект; я, разумеется, ни в коем случае не утверждаю, что данная

<sup>109</sup> В.Э.: Выше я уже говорил, что аналоговые модели не отличаются принципиально от дискретных моделей – если опуститься до достаточно низкого уровня, аналоговая модель всё равно оказывается дискретной.

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

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

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

### §1.21. Является ли невычислимым математическое воображение?

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

Дело в том, что «предметом» рассматриваемой выше «визуализации» является «визуальное» в буквальном смысле этого слова, т.е. мысленные образы, соответствующие, как нам представляется, сигналам, поступающим в мозг от глаз. В общем же случае мысленные образы вовсе не обязательно носят такой буквально «визуальный» характер – например, те, что возникают, когда мы понимаем смысл какого-то абстрактного слова или припоминаем музыкальную фразу. Согласитесь, что мысленные образы человека, слепого от рождения, вряд ли могут иметь прямое отношение к сигналам, которые его мозг получает от глаз. Иными словами, под «визуализацией» мы будем в дальнейшем подразумевать скорее процессы, связанные с «осознанием» вообще,<sup>110</sup> нежели те, что имеют непосредственное отношение к системе органов зрения. Честно говоря, мне не известен ни один довод, непосредственно указывающий на вычислительную (или какую-либо иную) природу нашей способности к визуализации именно в буквальном смысле этого слова. Моя же убежденность в том, что процессы «буквальной» визуализации действительно являются невычислимыми,<sup>111</sup> проистекает из явно невычислительного характера других видов осознания. Не совсем понятно, каким образом можно произвести прямое доказательство невычислимости исключительно для геометрической визуализации, однако если бы удалось убедительно доказать невычислимость хотя бы некоторых форм осмысленного осознания, то такое доказательство дало бы, по меньшей мере, серьезные основания полагать, что вид осознания, ответственный за геометрическую визуализацию, также должен иметь невычислительный характер. По-видимому, нет особой необходимости проводить четкую границу между различными проявлениями феномена сознательного понимания.

---

<sup>110</sup> В.Э.: Короче говоря, под «визуализацией» будут пониматься вообще все внутримозговые (внутрикомпьютерные) структуры данных, которые строят мозговые программы.

<sup>111</sup> В.Э.: Пенроуз всё время эту «вычислимость» и «невычислимость» понимает в каком-то «извращенном» смысле, – и я никак не могу до конца понять, что же он при этом имеет в виду. В чем проблема? Какие «невычислимые» «визуализации», когда мы в своих компьютерах каждый день видим и цветные картинки, и смотрим кинофильмы? О чем он говорит?

Переходя от общего к частному, я утверждаю, что наше понимание, например, свойств натуральных чисел (0, 1, 2, 3, 4,...) носит явно невычислительный характер. (Можно даже сказать, что само понятие натурального числа и есть, в некотором смысле, форма негеометрической «визуализации».)<sup>112</sup> В §2.5, воспользовавшись упрощенным вариантом теоремы Гёделя (см. пояснение к возражению Q16), я покажу, что это понимание невозможно описать каким бы то ни было конечным набором правил, а значит, невозможно и воспроизвести с помощью вычислительных методов. Время от времени нас радуют сообщениями о том, что ту или иную компьютерную систему «обучили» «пониманию» концепции натурального числа.<sup>113</sup> Однако, как мы вскоре увидим, этого просто не может быть. Именно осознание того, что в действительности может означать слово «число», дает нам возможность верно понять заключенную в нем идею. А располагая верным пониманием, мы – по крайней мере, в принципе – можем давать верные ответы на целый ряд вопросов о числах, буде нам таковые зададут, в то время как ни один конечный набор правил этого обеспечить не в состоянии. Имея в своем распоряжении одни только правила при полном отсутствии непосредственного осознания, управляемый компьютером робот (такой, например, как «*Deep Thought*»; см. §1.15) неизбежно окажется лишен тех способностей, в которых ни один из людей никаких ограничений не испытывает; хотя если снабдить робота достаточно умными правилами поведения, то он, возможно, поразит наше воображение выдающимися интеллектуальными подвигами, многие из которых далеко превзойдут способности обычного человека в каких-то конкретных, достаточно узкоспециальных областях. Возможно даже, что ему удастся на некоторое время одурачить нас, и мы поверим, что и он способен на осознание.

Следует отметить, что всякий раз, как мы получаем действительно эффективную цифровую (или аналоговую) компьютерную модель какой-либо внешней системы, это почти всегда происходит благодаря глубокому пониманию человеком тех или иных основополагающих математических идей. Взять хотя бы цифровую модель геометрического движения твердого тела. Выполняемые при таком моделировании вычисления опираются, главным образом, на открытия великих мыслителей семнадцатого века – таких, например, как французские математики Декарт, Ферма и Дезарг, – которым мы обязаны идеями системы координат и проективной геометрии. Существуют и модели, описывающие движение куска веревки или струны. Как выясняется, геометрические идеи, необходимые для понимания особенностей поведения струны – ее так называемой «заузленности», – весьма сложны и относительно молоды. Большинство фундаментальных открытий в этой области были сделаны только в двадцатом веке. Каждый из нас без особого труда способен экспериментальным путем – т.е. посредством несложных манипуляций руками и приложения некоторого здравого смысла – убедиться в наличии либо отсутствии на замкнутой, но спутанной веревочной петле узлов; вычислительные же алгоритмы для достижения того же результата оказываются на удивление сложными и малоэффективными.<sup>114</sup>

Таким образом, эффективное цифровое моделирование таких процессов является в основе своей нисходящим и во многом определяется пониманием и интуитивными прозрениями человека. Вероятность того, что в человеческом мозге при визуализации происходит нечто подобное, очень и очень невелика. Более правдоподобным представляется предположение о том, что

---

<sup>112</sup> В.Э.: На этом примере особенно ярко видно, как перекошены взгляды Пенроуза по сравнению с действительным положением вещей, поэтому зафиксируем и подчеркнем это еще раз! В представлениях Пенроуза числа – это «форма негеометрической визуализации» (в более точной модели, предоставленной Веданской теорией, это таксоны классификации множеств, проведенной программой N); далее у Пенроуза «понимание свойств натуральных чисел носит явно невычислительный характер» (эта «явная невычислимость», видимо, проистекает из того, что все свойства чисел невозможно до конца полно и непротиворечиво описать, как это следует из теоремы Гёделя) (на самом деле всё обстоит так: другие мозговые программы исследуют программу N, исследуют ее потенциальные продукты – числа; до конца познать эти свойства, конечно, не удастся, но некоторые свойства устанавливаются; во всем этом вообще нет ничего другого, кроме вычислительных процессов: исследуемый объект – это программа N, исследующий субъект – это соседние программы мозга). Остановитесь, читатель, и зафиксируйте у себя максимально четко эти два мировоззрения – Пенроузовское и Веданское – они крайне важны для понимания как самой этой книги, так и моих комментариев к ней!

<sup>113</sup> См., например, рассказ Фридмена [124] о работе Лената и других исследователей в этом направлении.

<sup>114</sup> В.Э.: Не те алгоритмы использовали. Человек использует другие алгоритмы, их и надо было брать для воспроизведения в компьютерах. (Только, видимо, не понимали, какие же именно алгоритмы использует человек).

существенный вклад в этот процесс вносят те или иные восходящие процедуры, а воспроизводимые в результате «визуальные образы» требуют предварительного накопления немалого «опыта». Я, впрочем, не слышал о сколько-нибудь серьезных исследованиях этого вопроса именно с точки зрения восходящих процедур (например, о разработках искусственных нейронных сетей). По всей видимости, подход, целиком основанный на процедурах восходящего типа, даст весьма скудные результаты. Сомневаюсь, что можно построить более или менее удачную модель геометрического движения твердого тела или топологических особенностей движения куска струны при отсутствии подлинного понимания обуславливающих эти движения законов.

Какие же физические процессы следует считать ответственными за осознание – за осознание, которое, судя по всему, необходимо для всякого подлинного понимания? Действительно ли оно не допускает численного моделирования, как того требует точка зрения  $\mathcal{C}$ ? Можно ли, в таком случае, надеяться на какое бы то ни было постижение этого предполагаемого физического процесса – хотя бы в принципе? Думаю, что можно, и более чем уверен, что точка зрения  $\mathcal{C}$  представляет собой подлинно научное допущение – просто нужно приготовиться к тому, что наши научные критерии и методы, возможно, претерпят не слишком заметные, но весьма существенные изменения. Нужно быть готовым к тому, что объекты наших исследований будут принимать самые неожиданные формы и возникать в таких областях подлинно научного знания, которые, на первый взгляд, никакого отношения к делу не имеют. Читателя, который намерен продолжить чтение этой книги, я прошу сохранять открытость восприятия и вместе с тем внимательно следить за рассуждениями и представляемыми научными свидетельствами, даже если они вдруг покажутся ему несколько сомнительными с точки зрения здравого смысла. Будьте готовы немного поразмыслить над предлагаемыми доводами,<sup>115</sup> а я, в свою очередь, приложу все усилия к изложению их в максимально доступном виде. Уверен, что, настроившись подобным образом, мы с вами преодолеем все преграды.

В оставшихся главах первой части я не буду касаться физики и возможных видов биологической активности, которые способны обусловить невычислимость, требуемую точкой зрения  $\mathcal{C}$ . Этими предметами мы займемся во второй части книги. Для начала нам предстоит решить вопрос об общей целесообразности поисков невычислимых процессов. Пока что вся целесообразность проистекает лишь из моей уверенности в том, что при сознательном понимании мы действительно выполняем какие-то невычислимые операции. Эту уверенность необходимо обосновать, для чего нам придется обратиться к математике.

### Послесловие к Первой главе

2010.08.13 13:20 пятница

**В.Э.:** Итак, я прочитал первую главу второй книги Пенроуза. Я комментировал ее, как и было задумано, по ходу чтения, в общем-то не зная, что будет написано дальше. Лишь однажды, прочитав несколько страниц вперед, вернулся назад и дописал там комментарий...

В Предисловии я поставил себе задачу: определить, в чем именно сбился Пенроуз, почему он приходит к таким (нелепым) выводам. Так в чем же он сбился?

Я думаю, что самое главное, что подвело Пенроуза – это «математическая парадигма». Это весь образ мышления, характерный для (традиционной) математики, но в первую очередь: «машины Тьюринга».

Пенроуз утверждает, что современные универсальные компьютеры эквивалентны «универсальной машине Тьюринга». Это я слышу всю жизнь. Уже в Университете преподаватели нам долбили: «ЭВМ эквивалентны машинам Тьюринга, ЭВМ эквивалентны машинам

---

<sup>115</sup> **В.Э.:** Всегда готов! – как пионер, – но только Пенроузу это ничего хорошего не сулит. Теорема Гёделя, сколь угодно тщательно разобранная, не докажет, что в мозге нет программы  $N$ , классифицирующей множества по количеству элементов, а «редукция волновой функции» (или что он там собирается предлагать для объяснения «сознания»), во-первых, не сможет объяснить конкретные явления психики (такие как сновидения, гипноз, истерия и др. – что запросто делает Веданская теория), а, во-вторых, Пенроуз не сможет обосновать, почему (в свете «лезвия Оккама») необходимы эти постулаты из области квантовой механики (да еще и только будущей, а пока не существующей квантовой механики!), если всё (и гораздо больше, чем при концепции Пенроуза) можно объяснить без этих постулатов, без новых законов природы, – при помощи обыкновенной, даже школьникам известной информатики.

Тьюринга...» А я уже тогда сомневался в этом. Но тогда меня это мало волновало: «Ладно, эквивалентны, так эквивалентны...»

Однако на самом деле машины Тьюринга НЕ ЭКВИВАЛЕНТНЫ компьютерам. Я не знаю, что они («теоретики» от алгоритмов) подразумевают под этой «эквивалентностью» и как ее доказывают. Возможно, в каком-то смысле она и существует, эта пресловутая «эквивалентность». Но в любом случае ее нет в главном.

Вы знаете, читатель, почему с 16-го века в Европе начался бурный расцвет математики – стали появляться все эти Кардано, Бомбелли, Вьеты, Декарты и прочие? Почему не раньше, не во времена Беды Благословенного?

Ответ меня в свое время поразил, когда я его впервые услышал: потому, что к этому времени в Европе повсеместно распространились «арабские» цифры и десятичная позиционная система счисления, которая полностью вытеснила «римские цифры» в математических расчетах! Стало возможно вычислять очень очень удобно, быстро, точно, глубоко, далеко – и тем самым познать законы чисел так, как это было совершенно невозможно при «римских цифрах», когда все вычислительные мощности мозга тратились на преодоление неудобств, созданных уродливой (по сравнению с теперешней) системой обозначения чисел.

«Теоретически» римские цифры эквивалентны арабским – и там и тут: есть число; оно как-то обозначается графически... А для практики они – НЕ эквивалентны!

И точно так же дело обстоит и с «машинами Тьюринга». Возможно, в каком-то смысле «чисто теоретически» они и эквивалентны компьютерам (как римские цифры арабским), но практически они представляют собой страшнейшие гири на ногах программистского мышления, цепи, сковывающие это мышление и не дающие ему достигнуть никаких достойных результатов, потому что вся «энергия мозга» уходит на преодоление неудобств, созданных этими «машинами». (Вот эти гири, эти цепи и погубили Пенроуза).

У меня давно, еще в 1980-е годы, укрепилось мнение: «Если человек рассуждает о машинах Тьюринга, значит, он профан в программировании и ничего стоящего не скажет».

Я расскажу вам немножко о своей операционной системе Диспос. Созданная во второй половине 1970-х, она была операционной системой для вычислительных сетей (которые тогда только стали создаваться). Машина, которой управлял мой Диспос, была (через адаптеры) связана с 15–20 другими машинами в нашем институте и соседних институтах Академии наук Латвии. К ней также были подключены два АЦПУ (печатающие устройства), два устройства ввода перфокарт,<sup>116</sup> несколько дисплеев и, конечно, множество внешних устройств дисков (такие шкафчики величиной как средняя стиральная машина).

Диспос вел обмен с соседними машинами, управлял своими внешними устройствами; ввод и вывод всегда велся параллельно с работой процессора. Запускаю, скажем, операцию вывода на АЦПУ строки, и компьютер работает дальше, допустим, запускает операцию вывода на дисплей, и опять работает дальше, допустим, запускает операцию обмена с соседней машиной, и опять работает дальше, скажем, запускает операцию вывода строки на втором АЦПУ, но тут приходит прерывание от первого АЦПУ, что строка распечатана; я обрабатываю это прерывание, разумеется, проверяю, не было ли ошибки и т.д., запускаю на этом АЦПУ следующую строку, тут приходит прерывание от дисплея, что вывод закончился, обрабатываю это прерывание, больше на этот дисплей выводить пока не надо, ставлю его в ожидание, тут приходит прерывание от другого дисплея: там оператор нажал «Ввод» и хочет что-то ввести, запускаю на нем операцию ввода, тут приходит прерывание от второго АЦПУ, что строка распечатана, запускаю следующую строку, тут приходит прерывание от другого адаптера: машина с соседнего института хочет мне что-то передать...

Я помню, однажды я зашел в машинный зал (кажется, просто для того, чтобы забрать свой листинг с АЦПУ). Там сидела оператор – замужняя женщина, имеющая маленького ребенка, – и что-то вязала, типа детских носков. Операторам было положено в определенное время запускать копирование дисков – обычные страховочные копии. Вот, она при мне оторвалась от своего вязания, запустила копирование и некоторое время понаблюдала за происходящим. Диски копируются, на операторском (главном) дисплее мелькают сообщения о скопированных файлах, а также о том, что такой-то машине отдан такой-то файл, что от такой-то машины получен такой-

<sup>116</sup> В течение 15-ти лет эксплуатации Диспоса конфигурация внешних устройств, конечно, многократно менялась. Не всегда были два устройства ввода карт или АЦПУ, временами обходились и одним и т.д., а в конце перфокарты вообще перестали использоваться, всё шло только через дисплеи.

то файл; в зале тем временем стоит сплошной грохот: два АЦПУ тарабандят одновременно, заканчивают один листинг, прогоняют бумагу и печатают следующий... Понаблюдала она за всем этим и потом обращается ко мне: «И как она не перепутывает, что ей делать?!»

В том-то и искусство программирования, чтобы не перепутать...

Вот так работает подлинная операционная система: параллельно ведет множество различных процессов, асинхронно реагирует на прерывания (внешние сигналы) – как на сигналы об окончании прежде мною запущенных операций, так и на инициативные – сигнализирующие о том, что кто-то хочет мне что-то передать. Все программы реентерабельны: мне всё равно, сколько будет АЦПУ – хоть сто: их обслуживает одна и та же программа; мне всё равно, сколько будет карточных вводов, сколько дисплеев, сколько дисков, сколько адаптеров – на каждый тип устройства своя реентерабельная программа и на каждое физическое устройство свой процесс и своя описывающая строчка при генерации Диспоса...

В Диспосе был гипервизор, который создавал виртуальную машину для ОС/360 (ОС/ЕС). Я мог работать на машине один, а мог запустить под собой ОС (стандартную операционную систему). Когда Диспос был на машине один, он занимал низшие адреса в памяти машины, а всё свободное место верхних адресов отводил на буфера (использовал как рабочую память). Но когда надо было запускать ОС, то на нижние адреса претендовал он, и он не мог запускаться в других адресах, потому что был отредактирован на определенное место памяти, и это обстоятельство изменить я не мог. Поэтому, когда загружался ОС, сам Диспос перебежал на другое место, в конец памяти, освобождая нижние адреса для ОС-а, а когда ОС снимался, перебежал обратно. Таким образом, Диспос был самоперемещающейся операционной системой, которая «бегала» по памяти туда-сюда...

Много еще всяких «штучек» было в моем Диспосе – обо всем не расскажешь. Но теперь представьте себе машину Тьюринга – и НА НЕЙ все эти «штучки»: множество реентерабельных программ, ведущих параллельно множество процессов и реагирующих на беспорядочные асинхронные прерывания, конечные и инициативные, на команды операторов и сигналы других машин, представьте резидентные и транзитные фазы этой системы, представьте систему, бегающую по памяти и запускающую под собой другую операционную систему...

Может быть Пенроуз может всё это представить на машине Тьюринга, но я не могу!

На самом деле, конечно, и Пенроуз этого не может – и никто не может. И поэтому всякий, кто рассуждает о машинах Тьюринга, тем самым отключает себя от сколь-нибудь профессионального программирования и ограничивает себя самыми примитивными алгоритмами, которые можно наглядно изобразить на «машинах Тьюринга». Ну, а дальше они делают вывод: «Ой, такими примитивными алгоритмами интеллект не построишь!»

Разумеется, не построишь – кто спорит!? Для интеллекта ой-ой-ой что нужно!

Я, вот, выше немножко обрисовал Диспос. Конечно, чтобы его запрограммировать и сделать, я должен был очень отчетливо и ясно представлять себе, КАК он должен быть устроен, КАК должны работать и взаимодействовать программы, чтобы в итоге они делали то, что я хочу.

Диспос мы можем обозначить как первый уровень профессионального программирования (недоступный уже машинам Тьюринга, т.е. этот уровень невозможно отобразить на машинах Тьюринга так, чтобы человек действительно понимал устройство системы и представлял его так, как представлял это я, когда создавал Диспос). Но на Диспосе действительный параллелизм существовал только между работой процессора и работой внешних устройств; в самом компьютере все процессы были лишь псевдопараллельны: пока один процесс работал, другие ждали.

Вторым уровнем профессионального программирования (еще менее доступным изображению на машинах Тьюринга) мы можем обозначить действительно параллельные процессы в многопроцессорной машине или, еще лучше, в сопряженных между собой машинах (как это и было при Диспосе, когда он через адаптеры соединялся с другими машинами). Такая параллельная (и асинхронная) работа многих процессоров – существенная составляющая интеллекта (как естественного, так и искусственного).

Но третий и высший уровень (уж еще менее изобразимый на машинах Тьюринга) – это манипуляции одних программ над другими программами. Это составляет ядро интеллекта – и столь (неудачно) искомый Пенроузом ключ к пониманию «сознания» и «разума». (Диспос, как и другие современные операционные системы, немножко манипулировал программами: создавал их загрузочные модули, транслируя их из ОС-овского листинга, загружал их, перемещал по памяти, но в целом, это, конечно, мелочи по сравнению с тем, КАКАЯ работа над программами

требуется для самопрограммирования – и, тем самым, для интеллекта. Тем не менее, я МОГУ себе представить системы всех трех названных уровней, и понимаю, КАК они должны быть устроены и работать).

А Пенроуз, сковав себя цепями «тьюринговых машин», отключился от всех высших уровней программирования, и поэтому НЕ МОЖЕТ представить себе, как должны работать и взаимодействовать программы интеллекта – естественного или искусственного. И поэтому он приходит туда, куда приходит.

Его классификация взглядов  $\mathcal{A}$ ,  $\mathcal{B}$ ,  $\mathcal{C}$  и  $\mathcal{D}$  не полна и не включает те взгляды, которые опираются на высшие уровни программирования и обозначены мною как  $\mathcal{E}$ .

Он рассматривает только имитации интеллекта (при помощи примитивных алгоритмов, изобразимых на машинах Тьюринга), но не рассматривает действительные реализации интеллекта при помощи высших уровней программирования.

Поэтому он не доказал ничего, что касалось бы Веданской теории.

Таков мой вердикт.

Я намереваюсь еще помещать в Векордию другие главы этой книги Пенроуза и его первую книгу (НРК), потому что мне это интересно, и, главное, разбирая примеры Пенроуза, весьма удобно показывать истинное положение дел согласно Веданской теории, но вынесенный здесь вердикт, я думаю, не изменится.

(Далее идет текст Пенроуза)

\* \* \*

## Глава 2. Гёделевское доказательство

### §2.1. Теорема Гёделя и машины Тьюринга

В наиболее чистом виде мыслительные процессы проявляются в сфере математики.<sup>117</sup> Если же мышление сводится к выполнению тех или иных вычислений, то математическое мышление, по всей видимости, должно обладать этим свойством в наибольшей степени. Однако, как это ни удивительно, в действительности всё происходит с точностью до наоборот. Именно математика дает нам самое явное свидетельство тому, что процессы сознательного мышления включают в себя нечто, не доступное вычислению.<sup>118</sup> Возможно, это покажется парадоксальным, однако для того, чтобы двигаться дальше, нам придется пока с этим парадоксом как-то примириться.

Прежде чем мы начнем, мне бы хотелось хоть как-то успокоить читателя в отношении математических формул, которые встретятся нам в нескольких последующих разделах (§§2.2–2.5), хотя надо признать, что страхи его не лишены оснований: ведь нам предстоит в какой-то мере уяснить для себя смысл и следствия ни много ни мало самой важной теоремы математической логики – знаменитой теоремы Курта Гёделя. Я привожу здесь очень и очень упрощенный вариант этой теоремы, опираясь, в частности, на несколько более поздние идеи Алана Тьюринга. Мы не будем пользоваться каким бы то ни было математическим формализмом, за исключением простейшей арифметики. Представленное доказательство, вероятно, будет кое-где несколько путаным, однако всего лишь путаным, а ни в коем случае не «сложным» в смысле необходимости каких-то предварительных познаний в математике. Воспринимайте доказательство в любом удобном для вас темпе и не стесняйтесь перечитывать его столько раз, сколько захочется. В дальнейшем (§§2.6–2.10) мы рассмотрим некоторые более специфические соображения, лежащие в основе теоремы Гёделя, однако читатель, не интересующийся подобными вопросами, может эти разделы пропустить без ущерба для понимания.

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

<sup>117</sup> В.Э.: Ну, это такое «математикоцентрическое» заявление... Во всяком случае, в математике мыслительные процессы, бесспорно, ПРОЯВЛЯЮТСЯ.

<sup>118</sup> В.Э.: Ничего подобного математика не дает, а дает именно противоположное. Сейчас увидим.

Среди положений, которые со всей неоспоримостью доказал Гёдель, имеется следующее: нельзя создать такую формальную систему логически обоснованных математических правил доказательства, которой было бы достаточно, хотя бы в принципе, для доказательства всех истинных теорем элементарной арифметики. Уже и это само по себе в высшей степени удивительно, однако это еще не всё. Многие говорят за то, что результаты Гёделя демонстрируют нечто большее, – а именно, доказывают, что способность человека к пониманию и постижению сути вещей невозможно свести к какому бы то ни было набору вычислительных правил. Иными словами, нельзя создать такую систему правил, которая оказалась бы достаточной для доказательства даже тех арифметических положений, истинность которых, в принципе, доступна для человека с его интуицией и способностью к пониманию, а это означает, что человеческие интуицию и понимание невозможно свести к какому бы то ни было набору правил.<sup>119</sup> Последующие мои рассуждения отчасти имеют целью убедить читателя в том, что вышеприведенное утверждение действительно следует из теоремы Гёделя; более того, именно на теореме Гёделя основывается мое доказательство неизбежности наличия в человеческом мышлении составляющей, которую никогда не удастся воспроизвести с помощью компьютера<sup>120</sup> (в том смысле, который мы вкладываем в этот термин сегодня).

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

Те из вас, кто читал мою предыдущую книгу «Новый разум короля» (см. НРК, глава 2), возможно, припомнят, что алгоритм там определяется как процедура, которую способна выполнить машина Тьюринга, или, если угодно, математически идеализированная вычислительная машина. Такая машина функционирует в пошаговом режиме, причем каждый ее шаг полностью задается нанесенной на рабочую «ленту» меткой, которую (метку) машина «считывает» в соответствующий момент времени, и «внутренним состоянием» машины (дискретно определенным) на этот момент. Количество различных разрешенных внутренних

---

<sup>119</sup> В.Э.: Здесь следует остановиться и осознать, какова же на самом деле ситуация. Итак, в человеческом мозге имеется программа N, порождающая натуральные числа (путем классификации множеств по количеству элементов) и имеется ряд других программ ( $M + \dots$ ), похожим способом порождающих другие основные понятия математики. Не зная о существовании и природе этих программ, Пенроуз те знания о их работе, которые у людей имеются, называет «математической интуицией» и «пониманием»; дальше эти «интуицию и понимание» (т.е. на самом деле программы  $N + M + \dots$ ) некоторые пытаются описать при помощи каких-то «формальных» правил (типа тех, что придумал Фреге); однако эти описания оказываются неполными и вообще правила не до конца пригодными (что доказывает Курт Гёдель); из этого Пенроуз (и компания) делают вывод, что «математическую интуицию» невозможно описать при помощи строгих правил и, значит, ее нельзя реализовать на компьютере. И вот здесь, понимая всё это, я не могу удержаться от смеха! Боже! КАКУЮ же «свинью» Пенроузу подложили его образ мышления и математическая парадигма! Ведь его так называемая «математическая интуиция» и есть на самом деле не что иное, как знания о (мозговой) программе! Программа сидит в самом начале всей этой цепочки, а Пенроуз с серьезной миной «доказывает», что ее не может быть! (Вспомнился конец бала у Сатаны, где мессир Воланд говорит Мастеру об Иване Бездомном: «Он едва самого меня не свел с ума, доказывая мне, что меня нету!») Короче: теорема Гёделя доказывает, что невозможно составить полные правила, для описания потенциальных продуктов программ  $N + M + \dots$ , но она НИЧЕГО не доказывает о существовании (или несуществовании) самих программ N, M и других в человеческой голове.

<sup>120</sup> В.Э.: Ну вот – все заблуждения Пенроуза как на ладони! Видно, откуда берется его убеждение в «невывисимости» сознания, и видно, ПОЧЕМУ оно ошибочно. Ведь нужно просто разделять эти три вещи: 1) что существует мозговая программа N, знания о которой Пенроуз именует «интуицией»; 2) что эту программу пытаются описать при помощи «формальных правил»; 3) что Гёдель доказывает безуспешность такой попытки (2), но не доказывает, что не существует программа (1) или что она не программа.

состояний конечно, общее число меток на ленте также должно быть конечным, хотя сама лента по длине не ограничена. Машина начинает работу с какого-то определенного состояния, которое мы обозначим, например, нулем «0», команды же подаются на ленте в виде, скажем, двоичного числа (т.е. последовательности нулей «0» и единиц «1»). Далее машина начинает считывать эти команды, передвигая ленту (либо, что то же самое, перемещаясь вдоль ленты) некоторым определенным образом, согласно встроенным пошаговым инструкциям, при этом действие машины на каждом этапе работы определяется ее внутренним состоянием и конкретным символом, считываемым на данном этапе с ленты. Руководствуясь всё теми же встроенными инструкциями, машина может стирать имеющиеся метки или ставить новые. В таком духе машина продолжает работать до тех пор, пока не достигнет особой команды «STOP», – именно в этот момент (и никак не раньше) машина прекращает работу, а мы можем увидеть на ленте ответ на выполнявшееся вычисление. Вот и всё, можно задавать машине новую задачу.

Можно представить себе некую особую машину Тьюринга, которая способна имитировать действие любой возможной машины Тьюринга. Такие машины Тьюринга называют универсальными. Иными словами, любая отдельно взятая универсальная машина Тьюринга оказывается в состоянии выполнить любое вычисление (или алгоритм), какое нам только может прийти в голову. Хотя внутреннее устройство современного компьютера весьма отличается от устройства описанной выше конструкции (а его внутренняя «рабочая область», пусть и очень велика, всё же не бесконечна, в отличие от идеализированной ленты машины Тьюринга), все современные универсальные компьютеры представляют собой, в сущности, универсальные машины Тьюринга.<sup>121</sup>

## §2.2. Вычисления

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

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

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

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

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

0, 1, 4, 9, 16, 25, 36, ...;

представленные в этом ряду числа получены следующим образом:

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

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

---

<sup>121</sup> В.Э.: Уже со студенческих времен я не могу понять, почему все «теоретики от алгоритмов» так держатся за эти «машины Тьюринга». Ну, предложил 23-летний юноша (только что окончивший университет и ставший «research fellow-ом» или, по-нашему, «научным сотрудником») Алан Тьюринг такую модель в мае 1936 года, когда в мире не существовало еще ни одного компьютера. Ну, было по тем временам это, пожалуй, действительно выдающимся достижением. Но с тех пор прошло 74 года – целая человеческая жизнь! Компьютеры теперь есть в каждом доме, с ними знаком почти каждый ребенок, основы компьютерного программирования учат в школе на уроках информатики! Если, как утверждает Пенроуз, машины Тьюринга эквивалентны компьютерам, то почему бы всё, что он хочет нам сказать об алгоритмах и процедурах, не продемонстрировать на примере настоящих (пусть идеализированных – в смысле бесконечной памяти и т.д.) компьютеров? Почему Пенроузу надо лезть в это беспросветное болото «тьюринговых машин», в которое ни один уважающий себя компьютерный программист за ним не последует? ПОЧЕМУ? Пенроуз не знает компьютеров? Или на реальных компьютерах все эти рассуждения не получаются? «Теоремы» не выходят? Они получаются только в Тьюринговых болотах?

$$\begin{array}{ccccccc} & & & & * & * & * & * \\ & & & * & * & * & & \\ & * & * & & * & * & * & * \\ , & * & , & * & * & , & * & * & * & * & , & \dots \\ & & & * & * & & * & * & * & * & \\ & & & * & * & & * & * & * & * & \end{array}$$

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

$$\begin{array}{cccccc} 0^2+0^2+0^2 & 0^2+0^2+1^2 & 0^2+0^2+2^2 & 0^2+1^2+1^2 & 0^2+1^2+2^2 \\ 0^2+2^2+2^2 & 1^2+1^2+1^2 & 1^2+1^2+2^2 & 1^2+2^2+1^2 & 2^2+2^2+2^2 \end{array}$$

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

### §2.3. Незавершающиеся вычисления

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

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

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

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

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

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

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

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

1, 3, 5, 7, 9, 11, 13, 15, 17, ....

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

(D) Найти четное число, большее 2, не являющееся суммой двух простых чисел. Вспомним, что простым называется натуральное число (отличное от 0 и 1), которое делится без остатка лишь само на себя и на единицу; иными словами, простые числа составляют следующий ряд:

2, 3, 5, 7, 11, 13, 17, 19, 23, ....

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

## §2.4. Как убедиться в невозможности завершить вычисление?

Мы установили, что вычисления могут как успешно завершаться, так и вообще не иметь конца. Более того, в тех случаях, когда вычисление завершиться в принципе не может, это его свойство иногда оказывается очевидным, иногда не совсем очевидным, а иногда настолько неочевидным, что ни у кого до сих пор не достало сообразительности однозначно такую невозможность доказать. С помощью каких методов математики убеждают самих себя и всех остальных в том, что такое-то вычисление не может завершиться? Применяют ли они при решении подобных задач какие-либо вычислительные (или алгоритмические) процедуры? Прежде чем мы приступим к поиску ответа на этот вопрос, рассмотрим еще один пример. Он несколько менее очевиден, чем (C), но все же гораздо проще (B). Возможно, нам удастся попутно получить некоторое представление о том, с помощью каких средств и методов математики приходят к своим выводам. В предлагаемом примере участвуют числа, называемые шестиугольными:

1, 7, 19, 37, 61, 91, 127, ....

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

```

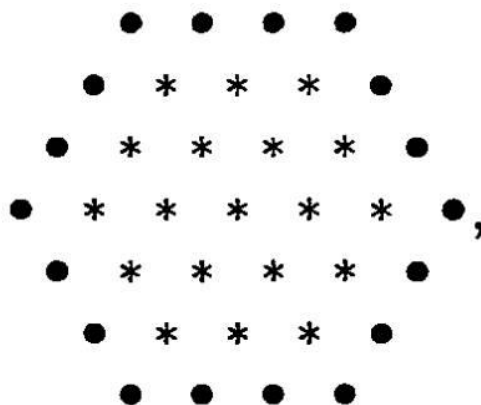
                                *  *  *  *
                                *  *  *  *  *
                                *  *  *  *  *
                                *  *  *  *  *  *
*,  *  *  *,  *  *  *  *  *,  *  *  *  *  *  *  *,  ....
                                *  *  *  *  *  *
                                *  *  *  *  *
                                *  *  *  *  *
                                *  *  *  *

```

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

6, 12, 18, 24, 30, 36, ....

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



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

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

$$1 = 1, \quad 1 + 7 = 8, \quad 1 + 7 + 19 = 27, \\ 1 + 7 + 19 + 37 = 64, \quad 1 + 7 + 19 + 37 + 61 = 125.$$

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

$$1 = 1^3 = 1 \times 1 \times 1, \quad 8 = 2^3 = 2 \times 2 \times 2, \quad 27 = 3^3 = 3 \times 3 \times 3, \\ 64 = 4^3 = 4 \times 4 \times 4, \quad 125 = 5^3 = 5 \times 5 \times 5, \dots$$

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

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

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

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

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

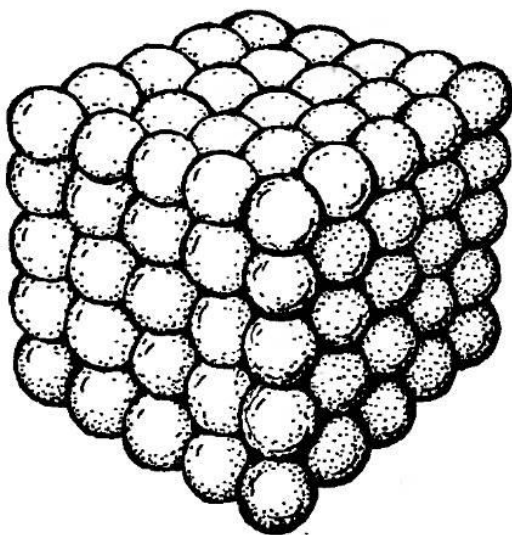


Рис. 2.1. Сферы, уложенные в кубический массив.

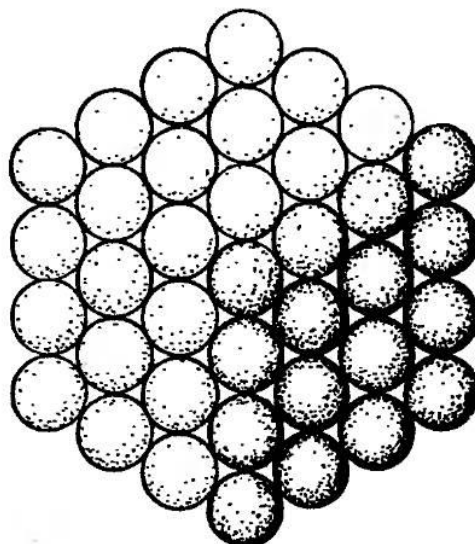


Рис. 2.3. Каждую часть построения можно рассматривать как шестиугольник.

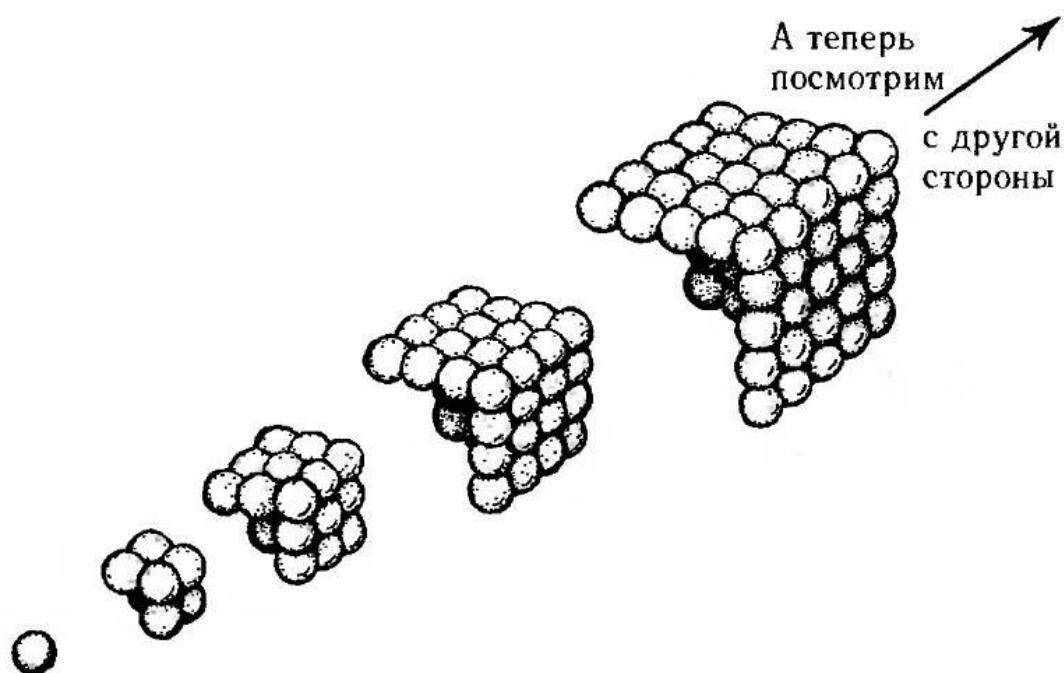


Рис. 2.2. Разберем куб на части – каждая со своей задней стенкой, боковой стенкой и потолком.

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

Посмотрим теперь на одну из наших трехгранных конфигураций со стороны, т.е. вдоль прямой, соединяющей начальную точку построения и точку, общую для всех трех граней. Мы увидим шестиугольник, подобный тому, что изображен на рис. 2.3. Точки, из которых складываются эти увеличивающиеся в размере шестиугольники, представляют собой, в сущности, те же точки, что образуют полный куб. То есть получается, что последовательное сложение шестиугольных чисел, начиная с единицы, всегда будет давать число кубическое. Следовательно, можно считать доказанным, что вычисление, требуемое для решения задачи (Е), никогда не завершится.

Кто-то, быть может, уже готов упрекнуть меня в том, что представленные выше рассуждения можно счесть в лучшем случае интуитивным умозаключением, но не формальным и строгим математическим доказательством. На самом же деле, перед вами именно доказательство, и доказательство вполне здоровое, а пишу всё это я отчасти и для того, чтобы показать, что осмысленность того или иного метода математического обоснования никак не связана с его «формализованностью» в соответствии с какой-либо заранее заданной и общепринятой системой правил.<sup>122</sup> Напомню, кстати, о еще более элементарном примере геометрического обоснования,

<sup>122</sup> В.Э.: Хорошо, а теперь разберем, что же здесь на самом деле произошло. Здесь отработал человеческий мозг (сначала Пенроуза, потом мой, потом Ваш, читатель, и т.д.) и пришел к определенным выводам. Посмотрим, какая работа этим мозгом была при этом выполнена по существу задачи (отбрасывая несущественные уклоны, какие, возможно, имели место). Итак, была дана задача (Е) «найти... сумму...» и была составлена мозговая программа *Е* для ее решения путем последовательного перебора подряд всех сумм «шестиугольных» чисел. Однако эта программа *Е* на выполнение не пускалась, а вместо этого была поставлена задача (Ж): определить, завершится ли программа *Е* вообще когда-нибудь, если ее пустить на выполнение? (Это типичная задача при самопрограммировании, т.е. при «разуме»: одни программы анализируются другими программами на предмет возможных последствий их выполнения). Для решения задачи (Ж) была составлена (и выполнена) мозговая программа *Ж*. В свою работу программа *Ж* вовлекала целый ряд других мозговых программ (либо существовавших уже раньше, либо созданных по ходу решения задачи (Ж)). Это программа *З*, складывающая числа  $(1 + 7 + 19 + 37 + 61 + 91 = 216 = 6 \times 6 \times 6 =$

применяемого для получения одного общего свойства натуральных чисел, – речь идет о доказательстве истинности равенства  $a \times b = b \times a$ , приведенном в § 1.19. Тоже вполне достойное «доказательство», хотя формальным его назвать нельзя.

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

Всегда ли для установления факта действительной незавершаемости вычисления достаточно применить некие четко определенные правила – такие, например, как принцип математической индукции? Как ни странно, нет. Это утверждение, как мы вскоре увидим, является одним из следствий теоремы Гёделя, и для нас крайне важно попытаться его правильно понять. Причем недостаточной оказывается не только математическая индукция. Недостаточным будет какой угодно набор правил, если под «набором правил» подразумевать некую систему формализованных процедур, в рамках которой возможно исключительно вычислительным путем проверить корректность применения этих правил в каждом конкретном случае. Такой вывод может показаться чересчур пессимистичным, ибо он, по-видимому, означает, что, несмотря на то, что вычисления, которые нельзя завершить, существуют, сам факт их незавершаемости строго математически установить невозможно. Однако смысл упомянутого следствия из теоремы Гёделя заключается вовсе не в этом. На самом деле, всё не так уж и плохо: способность понимать и делать выводы, присущая математикам – как, впрочем, и всем остальным людям, наделенным логическим мышлением и воображением, – просто-напросто не поддается формализации в виде того или иного набора правил.<sup>123</sup> Иногда правила могут стать частичной заменой пониманию, однако в полной мере такая замена не представляется возможной.

---

<sup>63</sup> и т.д.); это программа *И*, «визуализирующая» рост шестиугольников в пронумерованных Пенроузом рисунках этого параграфа; это программа *Й*, «визуализирующая» куб на рис.2.1, его рост на рис.2.2 и его проекцию на рис.2.3. (Отметим, что «визуализация» – это опять создание мозговой программы и анализ ее потенциальных продуктов со стороны. Так, например, программа *И* включает две подпрограммы: *И*<sub>1</sub>, которая строит ряд нарастающих шестиугольников, какие видны на первом нумерованном рисунке; и программа *И*<sub>2</sub>, которая, не выполняя реально *И*<sub>1</sub>, строит ее потенциальные продукты как визуальные образы, т.е. как структуры данных в памяти компьютера-мозга. Аналогично программа *Й* включает подпрограмму *Й*<sub>1</sub> построения видимых на рис.2.2 фигур, и подпрограмму *Й*<sub>2</sub> построения ее потенциальных продуктов в виде внутримозговых структур данных). И вот, программа *Ж* просматривает (анализирует) программы *З*, *И*, *Й* и их потенциальные продукты (предоставленные ей в виде внутримозговых структур данных) и находит связь между ними; она констатирует, что очередная прибавка в ряду рисунка 2.2 будет эквивалентна прибавке очередного шестиугольника в первом безымянном рисунке («эквивалентна» – это значит: программа *Н* отнесет их к одному и тому же таксону), и суммы шестиугольников будут эквивалентны кубам, и всё это будет соответствовать исчислениям программы *З*. А программа *Е* никогда не завершится, если ее пустить на выполнение. Вот та работа (кратко обозначенная), которую проделал в примере Пенроуза компьютер-мозг; при необходимости можно каждый блок разбирать подробнее и реализовывать на промышленных компьютерах. Пенроуз будет доказывать, что эта работа «сознательна», требует «понимания» и поэтому «невыводима», а у нас эту работу делает именно компьютер (хотя она действительно «сознательна» и требует «понимания»: и именно ТАКАЯ работа и скрывается за словами «сознательно», «понимание»).

<sup>123</sup> **В.Э.:** Задумаемся над этими словами Пенроуза. В примере, который мы только что разобрали, упоминаемая Пенроузом «способность понимать и делать выводы», осуществляется программой *Ж* при использовании программ *З*, *И*, *Й* и их подпрограмм. Разумеется, эти программы работают по каким-то алгоритмам, и эти алгоритмы можно и описать, и выполнить на других устройствах (технических компьютерах). Что же тогда такое «формализация в виде того или иного набора правил», которую сделать нельзя? Очевидно, это значит создать какие-то другие «правила» и алгоритмы, отличные от тех, по которым работают программы *Ж*, *З*, *И*, *Й*, и тогда на основе ИХ (этих правил) создать программу *К*,

## §2.5. Семейства вычислений; следствие Гёделя–Тьюринга $\mathcal{C}$

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

В терминах машин Тьюринга это всего лишь означает, что  $C(n)$  есть действие, производимое некоей машиной Тьюринга над числом  $n$ . Иными словами, число  $n$  наносится на ленту и подается на вход машины, после чего машина самостоятельно выполняет вычисления. Если вас почему-либо не устраивает концепция «машин Тьюринга», вообразите себе самый обыкновенный универсальный компьютер и считайте  $n$  «данными», необходимыми для работы какой-нибудь программы.<sup>124</sup> Нас в данном случае интересует лишь одно: при любом ли значении  $n$  может завершиться работа такого компьютера.

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

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

и

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

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

Предположим, что у нас имеется некая вычислительная процедура  $A$ , которая по своем завершении<sup>125</sup> дает нам исчерпывающее доказательство того, что вычисление  $C(n)$  действительно никогда не заканчивается. Ниже мы попробуем вообразить, что  $A$  включает в себя все известные математикам процедуры,<sup>126</sup> посредством которых можно убедительно доказать, что то или иное вычисление никогда не завершается. Соответственно, если в каком-то конкретном случае завершается процедура  $A$ , то мы получаем, в рамках доступного человеку знания, доказательство того, что рассматриваемое конкретное вычисление никогда не заканчивается. Большая часть последующих рассуждений не потребует участия процедуры  $A$  именно в такой роли, так как они посвящены, в основном, математическим умопостроениям. Однако для получения окончательного заключения  $\mathcal{C}$  нам придется-таки придать процедуре  $A$  соответствующий статус.

которая должна делать то же самое, что делает  $\mathcal{C}$ . (И тогда – по Гёделю–Пенроузу – окажется, что такую  $K$  нельзя создать? Интересно: а почему, собственно, нельзя?)

<sup>124</sup> В.Э.: Хорошо! Вот так и сделаем – и посмотрим, что тут выйдет. Итак  $C$  – это просто PROCEDURE, скажем, языка PASCAL, а на входе у нее один параметр:  $n$ . (И никаких машин Тьюринга!)

<sup>125</sup> Здесь я предполагаю, что если процедура  $A$  вообще завершается, то это свидетельствует об успешном установлении факта незавершаемости  $C(n)$ . Если же  $A$  «застревает» по какой-либо иной, нежели достижение «успеха», причине, то это означает, что в данном случае процедура  $A$  корректно завершиться не может. См. далее по тексту возражения Q3 и Q4, а также Приложение А, с. 193.

<sup>126</sup> В.Э.: В этом месте рассуждение Пенроуза, до сих пор достаточно ясное, теряет под собой почву реальности. Что это за процедура  $A$ , которая «включает в себя все известные математикам процедуры»? Я как программист привык рассуждать о таких программах, которые я (хотя бы в принципе) могу написать. Но эту  $A$  не может написать никто – ни я, ни кто другой. Поэтому рассуждения уже пошли о несуществующем объекте. Но, ладно – посмотрим, что будет дальше.

Я, разумеется, не требую, чтобы посредством процедуры  $A$  всегда можно было однозначно установить, что вычисление  $C(n)$  нельзя завершить (в случае, если это действительно так); однако я настаиваю на том, что неверных ответов  $A$  не дает, т.е. если мы с ее помощью пришли к выводу, что вычисление  $C(n)$  не завершается, значит, так оно и есть. Процедуру  $A$ , которая и в самом деле всегда дает верный ответ, мы будем называть обоснованной. Следует отметить, что если процедура  $A$  оказывается в действительности необоснованной, то этот факт, в принципе, можно установить с помощью прямого вычисления – иными словами, необоснованную процедуру  $A$  можно опровергнуть вычислительными методами. Так, если  $A$  ошибочно утверждает, что вычисление  $C(n)$  нельзя завершить, тогда как в действительности это не так, то выполнение самого вычисления  $C(n)$  в конечном счете приведет к опровержению  $A$ . (Возможность практического выполнения такого вычисления представляет собой отдельный вопрос, его мы рассмотрим в ответе на возражение Q8.)

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

$$C_0, C_1, C_2, C_3, C_4, C_5, \dots,^{127}$$

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

$$C_0(n), C_1(n), C_2(n), C_3(n), C_4(n), C_5(n), \dots^{128}$$

Можно представить, что эта последовательность задается, скажем, как некий пронумерованный ряд компьютерных программ. (Для большей ясности мы могли бы, при желании, рассматривать такую последовательность как ряд пронумерованных машин Тьюринга, описанных в НРК; в этом случае вычисление  $C_q(n)$  представляет собой процедуру, выполняемую  $q$ -й машиной Тьюринга  $T_q$  над числом  $n$ .) Здесь важно учитывать следующий технический момент: рассматриваемая последовательность является вычислимой – иными словами, существует одно-единственное<sup>129</sup> вычисление  $C^\bullet$ , которое, будучи выполнено над числом  $q$ , дает в результате  $C_q$ , или, если точнее, выполнение вычисления  $C^\bullet$  над парой чисел  $q, n$  (именно в таком порядке) дает в результате  $C_q(n)$ .<sup>130</sup>

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

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

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

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

Отметим, что  $A(n, n)$  зависит только от одного числа ( $n$ ), а не от двух, так что данное вычисление должно принадлежать ряду  $C_0, C_1, C_2, C_3, \dots$  (по  $n$ ), поскольку предполагается, что

<sup>127</sup> В.Э.: Так, стало быть, здесь он нумерует различные процедуры языка Паскаль, которые каждая делает свою работу: процедура  $C_0$ , процедура  $C_1$ , процедура  $C_2$  и т.д.

<sup>128</sup> В.Э.: То есть, всем этим процедурам подаем на вход одно и то же число.

<sup>129</sup> Собственно, точно такой же результат достигается посредством процедуры, выполняемой универсальной машиной Тьюринга над парой чисел  $q, n$ ; см. Приложение А и НРК, с. 51–57.

<sup>130</sup> В.Э.: Это сказано туманно, но, видимо, подразумевается, что если мы многократно обратимся к  $C_q$ , то это всегда будет одна и та же программа.

этот ряд содержит все вычисления, которые можно выполнить над одним натуральным числом  $n$ .<sup>131</sup> Обозначив это вычисление через  $C_k$ , запишем:

$$(J) \quad A(n, n) = C_k(n).$$

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

$$(K) \quad A(k, k) = C_k(k),$$

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

$$(L) \quad \text{Если завершается } A(k, k), \text{ то } C_k(k) \text{ не завершается.}$$

Подставляя (K) в (L), находим:

$$(M) \quad \text{Если завершается } C_k(k), \text{ то } C_k(k) \text{ не завершается.}^{132}$$

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

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

\* \* \*

2010.08.20 11:22 пятница

**В.Э.:** Сдержим эмоции, вспыхнувшие у меня и отраженные в последней сноске сразу после того, как я *осознал* и *понял* (любимые словечки Пенроуза) ЧТО ! мне пытаются всучить под видом математического доказательства. Разберем теперь это «доказательство» основательно, и для этого мне потребуется уже не подстрочное примечание, а крупная вставка в текст Пенроуза.

На рис. VE1 изображена схема, на которую лучше поглядывать, разбирая рассуждение Пенроуза, так как она дает некоторую «визуализацию» этих вещей.

По вертикальной оси в схеме отложены всевозможные вычислительные процедуры  $C_i(n)$ , имеющие на входе один параметр (число  $n$ ). Эти процедуры перенумерованы, и их общее количество –  $C$ .

По горизонтальной оси отложены числа, которые могут быть параметрами этих процедур; их общее количество –  $N$ .

Горизонтальные прерывистые линии показывают выполнение одной процедуры с всевозможными параметрами (входными числами). Вертикальные прерывистые линии показывают выполнение всевозможных процедур с одним и тем же параметром (числом). Пересечение горизонтальных и вертикальных прерывистых линий отображает выполнение одной определенной процедуры с одним определенным параметром (числом).

Но к каждому пункту этой плоскости привязана и другая процедура  $A(i, n)$  с двумя входными параметрами, первый из которых указывает на процедуру  $C_i$ , а второй – на ее входной параметр  $n$ . Эта процедура проверяет, заканчивается ли процедура  $C_i$  при данном параметре  $n$  (доказывает, будет она завершаться, или нет). Процедура  $A(i, n)$  такая хитрая, что ее можно рассматривать одновременно и как единую процедуру  $A$ , и как массив (размером  $C \times N$ ) отдельных разрозненных процедур, и как разные объединения (множества) этих разрозненных процедур (подмножества единой  $A$ ).

Таковы стартовые установки пенроузовского рассуждения, так сказать, «поле брани».

<sup>131</sup> **В.Э.:** Ах вот как! Оказывается, в том ряду  $C_0, C_1, C_2, C_3, \dots$  были все процедуры, которым можно подать на вход число  $n$ , в том числе и  $A$ ! (О! Теперь мы уже в двойном тумане!)

<sup>132</sup> **В.Э.:** Боже! И вот эту муру нам хотят преподнести как доказательство, имеющее какое-то отношение к чему-то в реальном мире! Во-первых, «процедура  $A$ », содержащая «все возможные доказательства», – это объект несуществующий; во-вторых, «ряд  $C$ », содержащий «все программы, которым можно подать на вход число  $n$ », – это объект несуществующий; в-третьих, даже если предположить, что  $A$  и  $C$  существуют, то нет никакой гарантии, что в пункте (I) можно будет взять  $q = n$ ; в-четвертых, программа  $A(n, n)$ , имеющая два параметра, не эквивалентна программе  $A(n)$ , имеющей один параметр; в-пятых, нет гарантии, что в пункте (K) можно будет взять  $n = k$ . Это то, что бросается в глаза сразу, с первого взгляда.

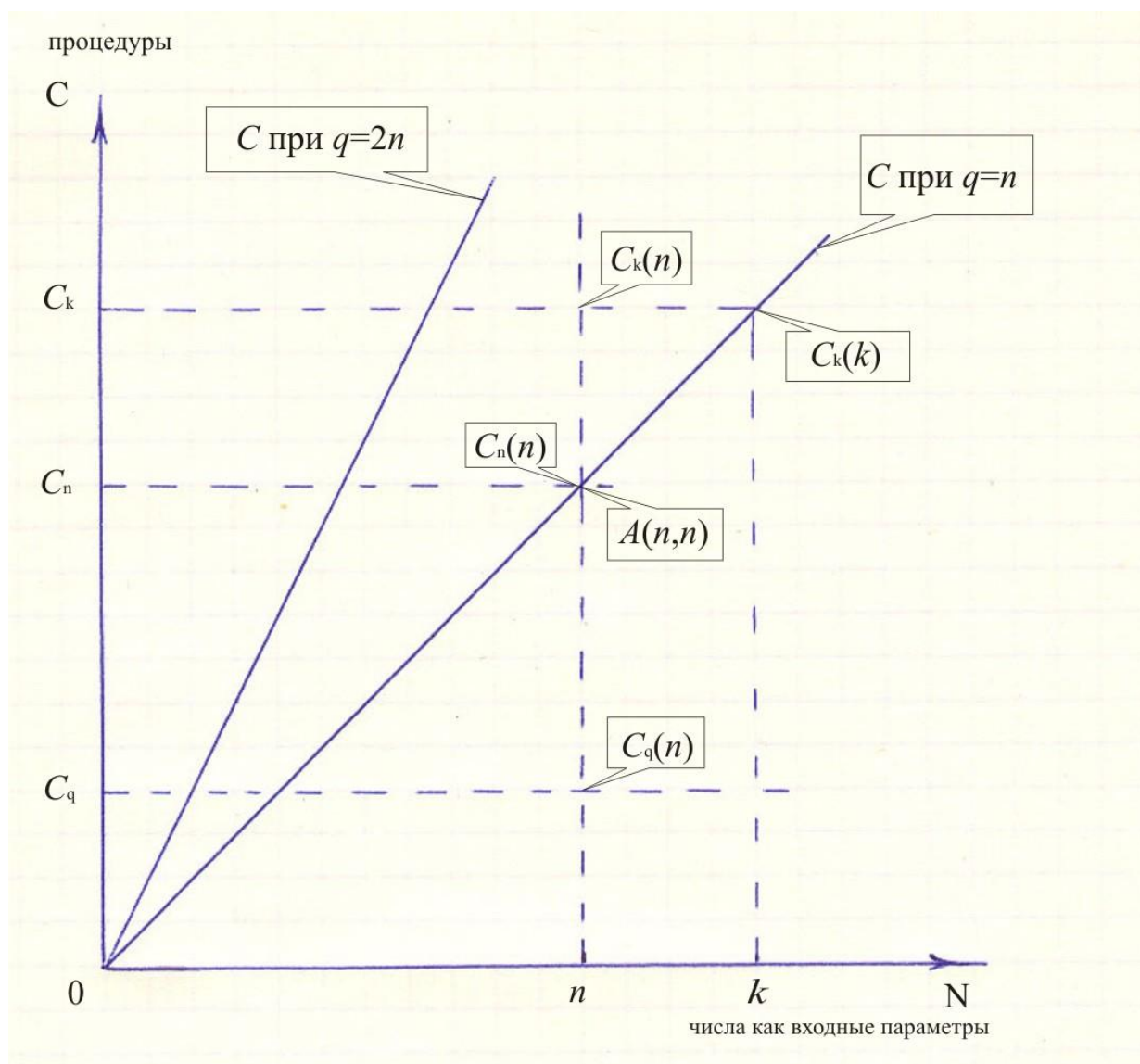


Рис.VE1. Схема для ориентации в рассуждении Пенроуза

Пенроуз перед своим доказательством дал нам разрешение: «Если вас почему-либо не устраивает концепция «машины Тьюринга», вообразите себе самый обыкновенный универсальный компьютер и считайте  $n$  «данными», необходимыми для работы какой-нибудь программы». Меня действительно «не устраивают» «машины Тьюринга», и поэтому я воспользуюсь разрешением Пенроуза и проверю его рассуждение на реальных программах реальных компьютеров, с которыми я имел достаточно много дел. Уж кто-кто, а я-то «обыкновенный компьютер» вообразить могу, как и программы с необходимыми для них данными!

Да только нет для нас, работающих программистов, такого ряда программ

$$C_0, C_1, C_2, C_3, \dots, C_q, \dots$$

Что я должен включать в этот ряд? Свою подпрограмму ABC, которая среди других входных параметров имеет  $n$ ? Или только те подпрограммы, которые имеют лишь один параметр –  $n$ ? А если процедура ABC сама с параметром  $n$  вообще не работает, а только передает его вызываемой функции DEF? Тогда что я должен включать в ряд  $C_q$  – только ABC? Или только DEF? Или обе программы? А если я вчера написал программу, а сегодня изменил ее? Тогда включать вчерашнюю или сегодняшнюю? А если программа ABC по алгоритму одинаковая, но написана для разных машин: для «Минск-22», для «Mitra-15», для «ЕС ЭВМ», для «IBM PC»? Тогда все включаем – или только одну? И если одну, то которую? А если моя программа и программа Леньки, моего друга по Институту, отличается только несколькими операторами и обе дают одинаковые результаты, то включаем обе или одну? И которую? А если я начал писать и не дописал программу, то включаем ее? Ведь как возможное вычисление она же существует! А ту,

которую я вообще не начал писать, а только подумал о ней? И ту, о которой подумал школьник, не умеющий программировать?.. Где граница?

Я программист, и привык мыслить точно и строго, и эта математическая расплывчатость и обычный для математиков туман мыслей меня никак не может устраивать. Если бы я рассуждал столь же туманно, как они, то мои программы не работали бы. Диспос перепутывал бы, что печатать на одном АЦПУ, что на другом, что выводить на дисплей, что копировать на другой диск, а что передавать другой машине. Всё пришло бы в сплошную кашу, застопорилось и зависло бы. Так что нас – программистов – проверяет самый придирчивый в мире судья: компьютер. Чуть что сделаешь не так – и всё полетит к чертям, работать не будет. Нас принуждают к абсолютно строгому мышлению – а кто на это не способен, тот не может быть программистом и вынужден уйти с этой работы.

Иное дело математики. Их никто не проверяет и не заставляет мыслить строго и точно. Придумал 23-летний Алан Тьюринг в 1936 году рассуждение, которое, как мы уже начали видеть и еще увидим дальше, представляет собой сплошную чушь, – и хорошо, и стал знаменитым. А в 1994 году уже и без того знаменитый 62-летний Роджер Пенроуз согласно кивает головой: «Да! Правильно! Всё верно! Гениально!». Сами себя проверили – и готово. Не требуется, чтобы что-то работало, действовало, не сбивалось и не зависало...

Я помню, во времена Диспоса однажды Ленька Рогов (один из лучших моих друзей в Институте электроники – он участвовал в других разработках, не Диспоса) увидел в журнале «Автоматика и вычислительная техника» статью Растрьгина (доктор ф.-м. наук, зав.лаб., главная знаменитость нашего Института после его директора академика Якубайтиса). В этой статье после преамбулы стояла начальная фраза:

«Возьмем программы  $P_1, P_2, P_3, \dots P_n$ »

И Ленька тогда сказал со своей обычной непревзойденной иронией:

– Вот, что значит доктор наук! Мы каждую программу пишем и отлаживаем неделями и месяцами, а он их просто берет охапками! – и закрыл журнал. (Да... все эти «теоретики от алгоритмов» для нас, пишущих и работающих программистов, были лишь объектами насмешек, и ни о каком авторитете их в наших глазах не могло быть и речи).

Итак, тот основной объект, над которым проводится рассуждение Пенроуза (множество процедур  $C_0, C_1, C_2, C_3, \dots, C_q, \dots$ ), для программиста не может считаться ни в малейшей мере определенным и осмысленным. Это всё равно что в детской сказке, когда мачеха приказывает падчерице: «Иди туда, не знаю, куда, возьми то, не знаю, что, а если до утра не принесешь...»

Не лучше обстоят дела и с процедурой  $A(i, n)$  – она тоже не представляет собой ничего такого, что мы, работающие программисты, могли бы воспринять как нечто реальное, хотя бы в принципе реализуемое на какой-нибудь машине и поэтому достойное обсуждения.

На самом деле одного этого уже достаточно, чтобы отвергнуть всё рассуждение Пенроуза как слишком туманное и не имеющее никакого отношения к реальному компьютерному программированию и к реальным программам.

Но изюминка еще впереди! Согласимся на время с установками Пенроуза и примем его «правила игры», чтобы посмотреть, что же из этого выйдет. Ладно, пусть у нас имеется ряд пронумерованных процедур  $C_0, C_1, C_2, C_3, \dots, C_q, \dots$  и хитроумная процедура  $A(i, n)$  для теоретической проверки, останавливаются ли все они при каждом  $n$  или нет.

Только выйдем всё-таки немножко из математического тумана и, во-первых, представим себе эти  $C_i$  более четко:

вот, допустим,  $C_0$  – это процедура возведения  $n$  в степень два (квадрат),

$C_1$  – это процедура извлечения квадратного корня из  $n$ ,

$C_2$  – это процедура извлечения кубического корня из  $n$  и т.д.,

а во-вторых, не будем сразу пускаться в туманную даль бесконечностей, а применим математическую индукцию.

Вот, выше, в §2.4 Пенроуз рисовал растущие шестиугольники, «визуализировал» добавление оболочек к кубам (рис. 2.2), и я полностью с ним соглашался. А потом Пенроуз писал:

«Представленное выше рассуждение о суммировании последовательных шестиугольных чисел можно при желании заменить более формальным математическим доказательством. В основу такого формального доказательства можно положить принцип математической индукции, т.е. процедуру установления истинности утверждения в отношении всех натуральных чисел на основании одного-единственного вычисления. По существу, этот принцип позволяет заключить, что

некое положение  $P(n)$ , зависящее от конкретного натурального числа  $n$  (например, такое: «сумма первых  $n$  шестиугольных чисел равна  $n^3$ »), справедливо для всех  $n$ , если мы можем показать, во-первых, что оно справедливо для  $n = 0$  (или, в нашем случае, для  $n = 1$ ), и, во-вторых, что из истинности  $P(n)$  следует истинность и  $P(n+1)$ ».

Вот, это правильно! И это действительно математика, действительно математическое мышление и действительно доказательство!

Поэтому сначала возьмем  $C$  (количество задействованных процедур) и  $N$  (количество возможных параметров) конечными и очень маленькими (ну, допустим,  $N = 4$ ), посмотрим, что получается при этих маленьких значениях, потом посмотрим, что получается при переходе к  $N + 1$ , и что получается, когда  $N \rightarrow \infty$ .

Итак, теперь в нашей «вселенной» имеется только пять чисел: 0, 1, 2, 3, 4.

Будет ли  $C$  (количество задействованных процедур) тоже ограничено этой же величиной (4)? Выше мы уже приняли, что процедуры  $C_0$ ,  $C_1$  и  $C_2$  у нас заняты (это возведение в квадрат, извлечение квадратного корня и кубического корня). Пока число  $C$  не превышает  $N$ . Но мы можем ввести сразу целых пять процедур умножения:  $C_3$  – умножить  $n$  на 0;  $C_4$  – умножить  $n$  на 1;  $C_5$  – умножить  $n$  на 2;  $C_6$  – умножить  $n$  на 3;  $C_7$  – умножить  $n$  на 4.

Теперь у нас количество процедур уже перевалило за  $N$  ( $N = 4$ , а  $C = 7$ ).

Следовательно, в диспозиции Пенроуза для  $q = 5$ ,  $q = 6$  и  $q = 7$  не будут существовать вычисления  $C_q(q)$ .

Пока еще это не очень страшно, так как на первом этапе «*диагонального доказательства – процедуры, открытой в высшей степени оригинальным и влиятельным датско-русско-немецким математиком девятнадцатого века Георгом Кантором*» Пенроуз «полагает  $q$  равным  $n$ ».

Это «положение» определяет диагональ, видную на рис. VE1. Только теперь у нас картина будет уже не квадратная, а вытянутая вверх ( $C = 7$ , а  $N = 4$ ). Поэтому диагональ упирается в правый край картины, не достигнув высоты  $C$ .

Разумеется, та пятерка процедур умножения на разные числа, которую мы добавили к  $C$ , не последняя. Мы можем добавить еще пятерку сложения  $n$  с разными числами, пятерку вычитания из  $n$  разных чисел, пятерку деления  $n$  на разные числа, и т.д. – вообще необозримое множество всевозможных процедур. Тогда количество  $C$  различных процедур будет еще возрастать, картина рисунка VE1 будет вытягиваться всё больше и больше вверх, а диагональ будет охватывать всё меньшую и меньшую часть этой картины.

Теперь возьмем  $N + 1 = 5$ . Что изменится? Теперь вместо пятерок процедур будут добавляться шестерки процедур, и растягивание картины станет еще быстрее.

Изменится ли что-нибудь при дальнейшем росте  $N$ ? Нет – не изменится. Картина как была, так и останется вытянутой, диагональ будет охватывать всё более и более ничтожную долю картины (при  $N \rightarrow \infty$  охваченная доля стремится к 0%).

Это вывод такой же достоверности, с какой в §2.4 Пенроуз рассуждал о растущих «шестиугольных числах» и кубах, при этом «визуализируя» процесс. Это действительно достоверный математический вывод.

А теперь перейдем к теперешнему рассуждению Пенроуза – к его второму этапу.

Итак, положили « $q$  равным  $n$ », получили  $C_n(n)$ ; заканчивается ли эта процедура, проверяет  $A(n,n)$ ; оно «зависит только от одного числа ( $n$ )» (оставим в стороне то обстоятельство, что для реальных компьютерных программ вообще-то есть разница, имеет ли процедура один параметр, или два параметра, значения которых оказались одинаковыми).

Теперь  $A(n,n)$  принадлежит к множеству  $C$ , имеет там номер  $k$ , ищем вычисление  $C_k(k)$ , оно должно находиться на пересечении горизонтали  $k$  с диагональю  $q = n$ , и... чушшш!.. эта горизонталь не пересекается с диагональю, потому что диагональ охватывает ничтожнейшую часть картины. Вычисление  $C_k(k)$  не найдено, оно не существует, никаких выводов сделать невозможно, доказательство лопнуло, получился пшик...

Вообще меня потрясает, как умные люди (а, по всему судя, Пенроуз же умный человек!) могут воспринимать всерьез «доказательства», основанные на «диагональном методе» Кантора. Ведь очевидно же, что этот метод противоречит математической индукции. Может быть достоверно лишь одно – либо математическая индукция, либо «диагональный метод». (И я без сомнений отвечаю, что достоверно первое, а второе – полная чушь). И то, что эта чушь теперь считается «общепринятой математической истиной», по-моему, говорит о глубоком вырождении математики. При Гауссе, когда математика еще была точной и достоверной наукой, такое «доказательство» было бы невозможно!)

«Доказательство», приводимое Пенроузом, состоит из сплошных изъянов. Там вообще нет ну абсолютно ничего, что могло бы заставить программиста сделать какие-то выводы относительно своих программ и вообще реального мира. Тем не менее Пенроуз на полном серьезе думает, что всё это действительно доказывает что-то о свойствах мозга, интеллекта, разума, сознания! Каким же должно быть мышление человека, чтобы ТАК полагать?!

Сам Пенроуз скажет, как только мы вернем ему слово: *«Надо признать, что, на первый взгляд, это доказательство и в самом деле смахивает на фокус»*. Нет, мистер Пенроуз! – не на фокус оно смахивает, а на манипуляции шамана дикого племени, продолжающего жить в каменном веке и верящего, что эти манипуляции представляют собой колдовство!

Знаменитый физик, лауреат Нобелевской премии 1965 года, Ричард Фейнман в своем выступлении перед выпускниками Калифорнийского технологического института в 1974 году (включенном в качестве Заключения в книгу «Вы, конечно же, шутите, мистер Фейнман...»<sup>133</sup>), говорил о псевдонауках, и там имеются такие слова:

«У тихоокеанских островитян есть религия самолетопоклонников. Во время войны они видели, как приземляются самолеты, полные всяких хороших вещей, и они хотят, чтобы так было и теперь. Поэтому они устроили что-то вроде взлетно-посадочных полос, по сторонам их разложили костры, построили деревянную хижину, в которой сидит человек с деревяшками в форме наушников на голове и бамбуковыми палочками, торчащими как антенны – он диспетчер, – и они ждут, когда прилетят самолеты. Они делают всё правильно. По форме всё верно. Всё выглядит так же, как и раньше, но всё это не действует. Самолеты не садятся. Я называю упомянутые науки науками самолетопоклонников, потому что люди, которые ими занимаются, следуют всем внешним правилам и формам научного исследования, но упускают что-то главное, так как самолеты не приземляются».



**Рис.VE2.** Вырезка из фильма режиссера Харальда Райнля (по книгам Эриха фон Дэнника) «Вспоминания о будущем» (1970). Самолетопоклонники выполняют ритуал призыва самолетов.

<sup>133</sup> См. {[R-FEYNMA](#)}.

Вот, это «доказательство», данное Пенроузом, – это всё равно что ритуальные манипуляции колдуна «самолетопоклонников»: они внешне подражают научному математическому доказательству, но на самом деле «упускают что-то главное», из-за чего «самолеты не приземляются», то есть, ничего на самом деле не доказано.

Но теперь, по крайней мере, я стал понимать, зачем «им» нужны машины Тьюринга. Совсем недавно, несколькими страницами выше, я так сокрушался и удивлялся:

«Уже со студенческих времен я не могу понять, почему все «теоретики от алгоритмов» так держатся за эти «машины Тьюринга». Ну, предложил 23-летний юноша (..) Алан Тьюринг такую модель в мае 1936 года, когда в мире не существовало еще ни одного компьютера. Ну, было по тем временам это, пожалуй, действительно выдающимся достижением. Но с тех пор прошло 74 года – целая человеческая жизнь! Компьютеры теперь есть в каждом доме...»

А теперь я понял! Машины Тьюринга «им» нужны для проведения «самолетопоклоннического» ритуала «доказательства» «проблемы остановки»! Это культовый предмет! С настоящими компьютерными программами ритуал не получается: нет даже видимости доказательства!..

Вообще с точки зрения программиста реальных компьютеров об этом рассуждении Пенроуза можно еще очень многое говорить. Что такое, например, «вычисление  $C_k$ » – то самое, которое при «пересечении с диагональю» даст легендарное  $C_k(k)$ ? Ведь  $C_k$  то же самое, что  $A(n, n)$  – оно «идет по диагонали» и проверяет:

- останавливается ли  $C_0$  при  $n = 0$ ?
- останавливается ли  $C_1$  при  $n = 1$ ?
- останавливается ли  $C_2$  при  $n = 3$ ?
- и т.д.

А ведь  $C_0$ ,  $C_1$ ,  $C_2$  – это вычисления совершенно различной природы! В наших, принятых выше, примерах, значит, процедура  $C_k$  будет работать так: если ей на вход подадут параметр 0, она проверит, можно ли возвести в квадрат число 0; если подадут на вход 1, то проверит, можно ли извлечь квадратный корень из 1; если подадут на вход 2, то проверит, можно ли извлечь кубический корень из 2; если подадут на вход 3, то проверит, можно ли умножить 4 на 0... Ну, и так далее: при каждом параметре происходят совершенно другие действия, не имеющие ничего общего с предыдущими! (Представляю проблемы того программиста, которому нужно написать эту программу  $C_k$ !).

Ладно,  $C_k$  идет «по диагонали». Но может ли быть какая-нибудь  $C_m$ , которая идет по более наклонной черте (рис. VE1) при  $q = 2n$ ? То есть, процедура  $A(2n, n)$  тоже является процедурой одного параметра и тоже представлена среди  $C$ ? (Во всяком случае в классической математике же функции  $y = x$  и  $y = 2x$  обе являются функциями одной переменной). А если так, то одних только процедур типа  $C_k$  будет в  $C$  столько же, сколько и натуральных чисел  $N$ .

Вообще очевидно, что мощность множества  $C$  бесконечно<sup>134</sup> раз больше, чем мощность множества  $N$ , и всё «доказательство» Пенроуза построено на игнорировании этого факта.

Но, пожалуй, хватит. Пойдем дальше.

Продолжим текст Пенроуза:

\* \* \*

В этом месте осторожный читатель, возможно, пожелает перечесть всё вышеприведенное доказательство заново, дабы убедиться в том, что он не пропустил какой-нибудь «ловкости рук» с моей стороны. Надо признать, что, на первый взгляд, это доказательство и в самом деле смахивает на фокус, и всё же оно полностью допустимо, а при более тщательном изучении лишь выигрывает в убедительности.<sup>135</sup> Мы обнаружили некое вычисление  $C_k(k)$ , которое, насколько нам известно, не завершается; однако установить этот факт с помощью имеющейся в нашем распоряжении вычислительной процедуры  $A$  мы не в состоянии. Это, собственно, и есть теорема Гёделя(–Тьюринга) в необходимом мне виде. Она применима к любой вычислительной процедуре  $A$ , предназначенной для установления невозможности завершить вычисление, – коль скоро нам известно, что упомянутая процедура обоснована. Можно заключить, что для однозначного установления факта незавершаемости вычисления не будет вполне достаточным ни один из

<sup>134</sup> Бесконечно? Или конечное число раз? Тут так сразу и не определишь из-за расплывчатости  $C$  – но очевидно, что мощность  $C$  во много много раз больше мощности  $N$ .

<sup>135</sup> В.Э.: Особенно после моего разбора ☺.

заведомо обоснованных наборов вычислительных правил (такой, например, как процедура  $A$ ), поскольку существуют незавершающиеся вычисления (например,  $C_k(k)$ ), на которые эти правила не распространяются. Более того, поскольку на основании того, что нам известно о процедуре  $A$  и об ее обоснованности, мы действительно можем составить вычисление  $C_k(k)$ , которое, очевидно, никогда не завершается, мы вправе заключить, что процедуру  $A$  никоим образом нельзя считать формализацией процедур, которыми располагают математики для установления факта незавершаемости вычисления, вне зависимости от конкретной природы  $A$ . Вывод:

$\mathcal{G}$  Для установления математической истины математики не применяют заведомо обоснованные алгоритмы.<sup>136</sup>

Мне представляется, что к такому выводу неизбежно должен прийти всякий логически рассуждающий человек. Однако многие до сих пор предпринимают попытки этот вывод опровергнуть (выдвигая возражения, обобщенные мною под номерами Q1 – Q20 в §2.6 и §2.10), и, разумеется, найдется ничуть не меньше желающих оспорить вывод более строгий, суть которого сводится к тому, что мыслительная деятельность непременно оказывается связана с некими феноменами, носящими фундаментально невычислительный характер.<sup>137</sup> Вы, возможно, уже спрашиваете себя, каким же это образом подобные математические рассуждения об абстрактной природе вычислений могут способствовать объяснению принципов функционирования человеческого мозга. Какое такое отношение имеет всё вышесказанное к проблеме осмысленного осознания? Дело в том, что, благодаря этим математическим рассуждениям, мы и впрямь можем прояснить для себя некие весьма важные аспекты такого свойства мышления, как понимание – в терминах общей вычислимости, – а, как было показано в §1.12, свойство понимания связано с осмысленным осознанием самым непосредственным образом. Предшествующее рассуждение действительно носит в основном математический характер, и связано это с необходимостью подчеркнуть одно очень существенное обстоятельство: алгоритм  $A$  участвует здесь на двух совершенно различных уровнях. С одной стороны, это просто некий алгоритм, обладающий определенными свойствами, с другой стороны, получается, что на самом-то деле  $A$  можно рассматривать как «алгоритм, которым пользуемся мы сами» в процессе установления факта незавершаемости того или иного вычисления. Так что в вышеприведенном рассуждении речь идет не только и не столько о вычислениях. Речь идет также и о том, каким образом мы используем нашу способность к осмысленному пониманию для составления заключения об истинности какого-либо математического утверждения – в данном случае утверждения о незавершаемости вычисления  $C_k(k)$ . Именно взаимодействие между двумя различными уровнями рассмотрения алгоритма  $A$  – в качестве гипотетического способа функционирования сознания и собственно вычисления – позволяет нам сделать вывод, выражающий фундаментальное противоречие между такой сознательной деятельностью и простым вычислением.

Существуют, однако, всевозможные лазейки и контраргументы, на которые необходимо обратить самое пристальное внимание. Для начала, в оставшейся части этой главы, я тщательно разберу все важные контраргументы против вывода  $\mathcal{G}$ ,<sup>138</sup> которые когда-либо попадались мне на глаза – см. возражения Q1–Q20 и комментарии к ним в §§2.6 и 2.10; там, кроме того, можно найти и несколько дополнительных возражений моего собственного изобретения. Каждое из возражений будет разобрано со всей обстоятельностью, на какую я только способен. Пройдя через это испытание, вывод  $\mathcal{G}$ , как мы убедимся, существенно не пострадает. Далее, в главе 3, я рассмотрю следствия уже из утверждения  $\mathcal{G}$ . Мы обнаружим, что оно и в самом деле способно

<sup>136</sup> В.Э.: Данное выше Пенроузом «доказательство» ни в малейшей мере не является таким рассуждением, которое мог бы воспринять всерьез программист, привыкший мыслить строго, точно и конкретно. Для такого программиста это «доказательство» не доказывает абсолютно ничего. Но, несмотря на это, тезис  $\mathcal{G}$  – верный. Просто он вытекает не из (туманного) рассуждения Пенроуза, а из других соображений. Если под «заведомо обоснованными алгоритмами» понимать «правила» и «доказательства» формальных систем (а Пенроуз понимает именно это, что особенно ясно станет ниже), то, действительно, НЕ на эти правила-алгоритмы формалистов опирается познание математических истин. А на какие алгоритмы опирается на самом деле – это показывает Веданская теория.

<sup>137</sup> В.Э.: Вот как тут Пенроуз перескакивает и сбивается! Для него понятия «правила-алгоритмы формалистов» и «вычислительные процессы» совпадают. Нет вторых вне первых. Поэтому из  $\mathcal{G}$  следует невычислительная природа мышления вообще. Но на самом деле понятия «правила-алгоритмы формалистов» и «вычислительные процессы» НЕ совпадают. На правилах-алгоритмах формалистов мышление действительно не опирается. Но оно представляет собой другие вычислительные процессы.

<sup>138</sup> В.Э.: Против вывода  $\mathcal{G}$ ? Значит, против тезиса, который я НЕ оспариваю.

послужить прочным фундаментом для построения весьма убедительного доказательства абсолютной невозможности точного моделирования сознательного математического понимания посредством вычислительных процедур,<sup>139</sup> будь то восходящих, нисходящих или любых их сочетаний. Многие сочтут такой вывод весьма неприятным, поскольку если он справедлив, то нам, получается, просто некуда двигаться дальше. Во второй части книги я выберу более позитивный курс. Я приведу правдоподобные, на мой взгляд, научные доводы в пользу справедливости результатов моих размышлений о физических процессах, которые могут, предположительно, лежать в основе деятельности мозга – вроде той, что осуществляется при нашем восприятии приведенных выше рассуждений, – и о причинах недоступности этой деятельности для какого бы то ни было вычислительного описания.

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

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

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

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

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

«первые 10 этапов  $A_1$ ;  
вторые 10 этапов  $A_1$ , первые 10 этапов  $A_2$ ;  
третьи 10 этапов  $A_1$ , вторые 10 этапов  $A_2$ , первые 10 этапов  $A_3$ ;  
... и т.д.»...

<sup>139</sup> В.Э.: Ну нет же: только с процедурами математических формалистов ты справишься, но не с компьютерами вообще.

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

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

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

Для описания изменяющегося алгоритма необходимо каким-то образом задать правила, согласно которым он, собственно, изменяется. Если сами по себе эти правила являются полностью алгоритмическими, то мы уже включили их в описание нашей гипотетической процедуры « $A$ », иначе говоря, такой «изменяющийся алгоритм» на деле представляет собой всего-навсего еще один пример единичного алгоритма, и на наши рассуждения подобное допущение никак не влияет. С другой стороны, можно вообразить средства для изменения алгоритма, предположительно не являющиеся алгоритмическими: такие, например, как введение в алгоритм каких-то случайных составляющих или неких процедур взаимодействия его с окружением. «Неалгоритмический» статус подобных средств изменения алгоритма мы еще будем рассматривать несколько позднее (см. §§ 3.9, 3.10); можно также вернуться к §1.9, где было показано, что ни одно из этих средств не позволяет сколько-нибудь убедительно избавиться от алгоритмизма<sup>140</sup> (как того требует точка зрения  $\mathcal{C}$ ). В данном случае, т.е. в рамках чисто математических рассуждений, нас занимает лишь возможность того, что такое изменение действительно будет носить алгоритмический характер. Если же предположить, что алгоритмическим оно быть никак не может, то мы, безусловно, придем к полному согласию с выводом  $\mathcal{C}$ .

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

$$A_0(q, n), A_1(q, n), A_2(q, n), A_3(q, n), \dots,$$

каждый элемент которого предположительно является обоснованной процедурой для установления незавершаемости вычисления  $C_q(n)$ , при этом мы будем считать, что мощность этих процедур возрастает по мере увеличения  $t$ . Предполагается также, что способ, посредством которого увеличивается мощность этих процедур, является алгоритмическим. Возможно, этот «алгоритмический способ» зависит некоторым образом от «опыта» выполнения предыдущих алгоритмов  $A_t(q, n)$ , однако в данном случае мы предполагаем, что этот «опыт» порождается также алгоритмически (в противном случае мы снова приходим к согласию с  $\mathcal{C}$ ), т.е. мы имеем полное право включить «опыт» (или способы его порождения) в перечень операций, составляющих следующий алгоритм (т.е., собственно, в  $A_t(q, n)$ ). Действуя таким образом, мы опять-таки получаем единичный алгоритм ( $A_t(q, n)$ ), который зависит алгоритмически от всех трех параметров:  $t, q, n$ . На его основе можно построить алгоритм  $A^*$ , столь же мощный, что и весь ряд  $A_t(q, n)$ , однако зависящий только от двух натуральных чисел:  $q$  и  $n$ . Для получения такого  $A^*(q, n)$  нам, как и прежде, необходимо лишь выполнить первые десять шагов алгоритма  $A_0(q, n)$  и запомнить результат; затем первые десять шагов алгоритма  $A_1(q, n)$  и вторые десять шагов алгоритма  $A_0(q, n)$ , запоминая получаемые результаты; затем первые десять шагов алгоритма  $A_2(q, n)$ , вторые десять шагов алгоритма  $A_1(q, n)$ , третьи десять шагов алгоритма  $A_0(q, n)$  и т.д., запоминая получаемые на каждом шаге вычисления результаты. В конечном итоге,

<sup>140</sup> Термин «алгоритмизм», который (по своей сути) прекрасно подходит для обозначения «точки зрения  $\mathcal{A}$ » в моей классификации, был предложен Хао Ваном [377].

сразу после завершения любого из составляющих алгоритм вычислений завершается выполнение и всей процедуры в целом. Замена процедуры  $A$  процедурой  $A^*$  никак не влияет на ход рассуждений, посредством которых мы пришли к выводу  $\mathcal{G}$ .

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

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

Если вас такое положение дел не устраивает, попробуйте представить алгоритм  $A$  следующим образом: пусть  $A$  объединяет в себе оба вида понимания, но в том случае, когда выясняется, что вычисление  $C_q(n)$  действительно завершается, алгоритм  $A$  искусственно заикливается (т.е. выполняет какую-то операцию снова и снова, бесконечное количество раз). Разумеется, на самом деле математики работают иначе, однако дело не в этом. Наше рассуждение построено как *reductio ad absurdum*,<sup>141</sup> т.е. начав с допущения, что для установления математической истины используются заведомо обоснованные алгоритмы, мы в итоге приходим к противоположному выводу. Такое доказательство не требует, чтобы гипотетическим алгоритмом непременно оказался какой-то конкретный алгоритм  $A$ , мы вполне можем заменить его на другой алгоритм, построенный на основе  $A$ , – как, например, в только что упомянутом случае.

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

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

В самом деле, было бы весьма затруднительно однозначно гарантировать, что каждому натуральному числу  $q$  в нашей последовательности действительно соответствует некое рабочее вычисление  $C_q$ . Например, описанная в НРК последовательность машин Тьюринга  $T_q$  этому условию, конечно же, не удовлетворяет; см. НРК, с. 54. При определенных значениях  $q$  машину Тьюринга  $T_q$  можно назвать «фиктивной» по одной из четырех причин: ее работа никогда не завершается; она оказывается «некорректно определенной», поскольку представление числа  $n$  в виде двоичной последовательности содержит слишком много (пять или более) единиц подряд и, как следствие, не имеет интерпретации в данной схеме; она получает команду, которая вводит ее в нигде не описанное внутреннее состояние; или же по завершении работы она оставляет ленту пустой, т.е. не дает никакого численно интерпретируемого результата. (См. также Приложение А.) Для приведенного в §2.5 доказательства Гёделя–Тьюринга вполне достаточно объединить все эти причины в одну категорию под названием «вычисление не завершается». В частности, когда я говорю, что вычислительная процедура  $A$  «завершается» (см. также примечание на с. 122)<sup>142</sup>, я подразумеваю, что она «завершается» как раз в вышеупомянутом смысле (а потому не содержит неинтерпретируемых последовательностей и не оставляет ленту пустой), – иными словами,

<sup>141</sup> Приведение к абсурду (лат.), доказательство от противного. *Прим. перев.*

<sup>142</sup> В.Э.: На с. 122 русского издания «Теней Разума» (этого) нет никакого примечания.

«завершиться» может только действительно корректно определенное рабочее вычисление. Аналогично, фраза «вычисление  $C_q(n)$  завершается» означает, что данное вычисление корректно завершается именно в этом смысле. При такой интерпретации соображение Q4 не имеет совершенно никакого отношения к представленному мною доказательству.

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

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

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

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

Такое рассуждение, основываясь на *reductio ad absurdum*, не просто показывает, что требуемому условию не соответствует некое частное простое число  $p$ , поскольку можно отыскать число больше него; оно показывает, что наибольшего простого числа просто не может существовать в природе. Аналогично, представленное выше доказательство Гёделя–Тьюринга не просто показывает, что нам не подходит тот или иной частный алгоритм  $A$ , оно демонстрирует, что в природе не существует алгоритма (познаваемо обоснованного), который был бы эквивалентен способности человека к интуитивному пониманию, которую мы применяем для установления факта незавершаемости тех или иных вычислений.<sup>143</sup>

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

Отыскание конкретного вычисления  $C_k(k)$  при заданном алгоритме  $A$ , безусловно, представляет собой вычислительный процесс. Более того, это можно достаточно явно показать.<sup>144</sup> Означает ли это, что предположительно неалгоритмическая математическая интуиция — интуиция, благодаря которой мы определяем, что вычисление  $C_k(k)$  никогда не завершается — на деле является всё же алгоритмической?

Думаю, данное суждение следует рассмотреть более подробно, поскольку оно представляет собой одно из наиболее распространенных недоразумений, связанных с гёделевским доказательством. Следует особо уяснить, что оно не сводит на нет ничего из сказанного ранее. Хотя процедуру отыскания вычисления  $C_k(k)$  с помощью алгоритма  $A$  можно представить в виде вычисления, это вычисление не входит в перечень процедур, содержащихся в  $A$ . И не может

<sup>143</sup> В.Э.: Доказательство Евклида опирается на алгоритм нахождения числа  $N = 2 \times 3 \times 5 \times \dots \times p + 1$ . Этот алгоритм выполним и дает результат. Рассуждение, которое Пенроуз называет «доказательством Гёделя–Тьюринга», опирается на алгоритм нахождения элемента (вычисления)  $C_k(k)$ . Но этот алгоритм не выполним в общем случае, так как не находится  $n$ , равное  $k$  (при «визуализации» в рис. VE1: горизонталь  $k$  не пересекается с диагональю  $q = n$ ).

<sup>144</sup> Чтобы подчеркнуть, что я принимаю это обстоятельство во внимание, я отсылаю читателя к Приложению А, где представлена явная вычислительная процедура (выполненная в соответствии с правилами, подробно описанными в НРК, глава 2) для получения операции  $C_k(k)$  машины Тьюринга посредством алгоритма  $A$ . Здесь предполагается, что алгоритм  $A$  задан в виде машины Тьюринга  $T_a$ , определение же вычисления  $C_q(n)$  кодируется как операция машины  $T_a$  над числом  $q$ , а затем над числом  $n$ .

входить, поскольку самостоятельно алгоритм  $A$  не способен установить истинность  $C_k(k)$ , тогда как новое вычисление (вкуче с  $A$ ), судя по всему, вполне на это способно. Таким образом, несмотря на то, что с помощью нового вычисления действительно можно отыскать вычисление  $C_k(k)$ , членом клуба «официальных установителей истины» оно не является.

Изложим всё это несколько иначе. Вообразите себе управляемого компьютером робота, способного устанавливать математические истины с помощью алгоритмических процедур, содержащихся в  $A$ . Для большей наглядности я буду пользоваться антропоморфной терминологией и говорить, что робот «знает» те математические истины (в данном случае – связанные с установлением факта незавершаемости вычислений), которые он может вывести, применяя алгоритм  $A$ . Однако если наш робот «знает» лишь  $A$ , то он никак не сможет «узнать», что вычисление  $C_k(k)$  не завершается, даже если процедура отыскания  $C_k(k)$  с помощью  $A$  является целиком и полностью алгоритмической. Мы, разумеется, могли бы сообщить роботу о том, что вычисление  $C_k(k)$  и в самом деле не завершается (воспользовавшись для установления этого факта собственными пониманием и интуицией), однако, если робот примет это утверждение на «веру», ему придется изменить свои собственные правила, присоединив полученную новую истину к тем, что он уже «знает». Мы можем пойти еще дальше и каким-либо способом сообщить нашему роботу о том, что для получения новых истин на основании старых ему, помимо прочего, необходимо «знать» и общую вычислительную процедуру отыскания  $C_k(k)$  посредством алгоритма  $A$ . К запасу «знаний» робота можно добавить всё, что является вполне определенным и вычислительным по своей природе. Однако в результате у нас появляется новый алгоритм « $A$ », и доказательство Гёделя следует применять уже к нему, а не к старому  $A$ . Иначе говоря, везде вместо старого  $A$  нам следовало бы использовать новый « $A$ », поскольку менять алгоритм « $A$ » посреди доказательства есть не что иное, как жульничество. Таким образом, как мы видим, изъясн возражения Q6 очень похож на рассмотренный выше изъясн Q5. В нашем *reductio ad absurdum* мы полагаем, что алгоритм  $A$  (под которым понимается некая познаваемая и обоснованная процедура для установления факта незавершаемости вычислений) в действительности представляет собой всю совокупность известных математикам подобных процедур, из чего и следует противоречие. Попытку введения еще одной вычислительной процедуры для установления истины – процедуры, не содержащейся в  $A$ , – после того как мы договорились, что  $A$  представляет собой всю их совокупность, я расцениваю как жульничество.

Беда нашего злосчастного робота в том, что, не обладая каким бы то ни было пониманием гёделевской процедуры, он не располагает ни одним надежным и независимым способом установления истины – истину ему сообщаем мы. (Эта проблема, вообще говоря, не имеет никакого отношения к вычислительным аспектам доказательства Гёделя.) Для того, чтобы достичь чего-то большего, ему, как и всем нам, необходимо понимание смысла операций, которые ему велено выполнять. Если такого понимания нет, то он вполне может «знать» (ошибочно), что вычисление  $C_k(k)$  завершается, а вовсе не наоборот. Заключение (ошибочное) «вычисление  $C_k(k)$  завершается» выводится точно так же алгоритмически, как и заключение (правильное) «вычисление  $C_k(k)$  не завершается». Таким образом, дело вовсе не в алгоритмическом характере этих операций, а в том, что для различения между алгоритмами, приводящими к истинным заключениям, и теми, что приводят к заключениям ложным, наш робот нуждается в способности выносить достоверные суждения об истинности. Далее, на данной стадии рассуждения, мы всё еще допускаем возможность того, что процесс «понимания» представляет собой некую разновидность алгоритмической деятельности, которая не содержится ни в одной из точно заданных и «заведомо» обоснованных процедур типа  $A$ . Например, понимание может осуществляться посредством выполнения какого-то необоснованного или непознаваемого алгоритма. В дальнейшем (см. главу 3) я попробую убедить читателя в том, что в действительности понимание вообще не является алгоритмической деятельностью. На настоящий же момент нас интересуют всего лишь строгие следствия из доказательства Гёделя–Тьюринга, а на них возможность получения вычисления  $C_k(k)$  из процедуры  $A$  вычислительным путем никоим образом не влияет.

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

Несмотря на кажущуюся логичность этого утверждения, здесь упущен из виду один очень существенный момент, а именно: способ, посредством которого мы (или компьютеры) определяем, какие математические утверждения истинны, а какие – ложны. (Во всяком случае, на простое хранение математических утверждений способны и системы, гораздо менее сложные, нежели универсальный компьютер – например, фотоаппараты.) Принцип использования компьютера в Q7 совершенно не учитывает критического вопроса о наличии у этого самого компьютера способности суждения об истинности. С равным успехом можно вообразить и компьютеры, в памяти которых не содержится ничего, кроме перечня абсолютно ложных математических «теорем», либо случайным образом перемешанных истинных и ложных утверждений. Откуда мы узнаем, какому компьютеру можно доверять? Я отнюдь не утверждаю, что эффективное моделирование результатов сознательной интеллектуальной деятельности человека (в данном случае, в области математики) абсолютно невозможно, поскольку по одной лишь чистой случайности компьютер может «умудриться» сделать всё правильно, пусть и не обладая каким бы то ни было пониманием. Однако шансы на это до абсурдного малы, в то время как те вопросы, на которые мы здесь пытаемся найти ответ (например, каким таким образом мы определяем, что вот это математическое утверждение истинно, а вот это – ложно?), в возражении Q7 и вовсе не затрагиваются. С другой стороны, Q7 всё же напоминает об одном более существенном соображении. Имеет ли непосредственное отношение к нашему исследованию обсуждение бесконечных структур (всех натуральных чисел или всех вычислений), если учесть, что совокупность всех результатов, полученных на тот или иной момент времени всеми людьми и компьютерами, имеет конечную величину? В следующем комментарии мы рассмотрим этот безусловно важный вопрос отдельно.

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

Всё верно, рассуждая в идеализированном ключе о машинах Тьюринга, незавершающихся вычислениях и т.п., мы рассматривали бесконечные (потенциально) процессы, тогда как в случае людей или компьютеров нам приходится иметь дело с системами конечными. И, разумеется, применяя подобные идеализированные доказательства к реальным и конечным физическим объектам, следует быть готовыми к тому, что такая операция непременно окажется связанной с теми или иными ограничениями и оговорками. Однако, как выясняется, учет конечной природы реальных объектов не изменяет сколько-нибудь существенно сути доказательства Гёделя–Тьюринга. Нет ничего странного в том, что мы рассуждаем об идеализированных вычислениях, обосновываем те или иные умозаключения и выводим, математически, их теоретические ограничения. Можно, к примеру, обсуждать в абсолютно конечных терминах вопрос о том, существует ли нечетное число, являющееся суммой двух четных чисел, или существует ли натуральное число, не являющееся суммой четырех квадратов (как в приведенных выше задачах (C) и (B)), нисколько не смущаясь тем, что при рассмотрении этих вопросов мы неявно учитываем бесконечное множество всех натуральных чисел. Мы имеем полное право рассуждать о незавершающихся вычислениях или машинах Тьюринга вообще, как о математических структурах, пусть и не в силах создать на практике бесконечно работающую машину Тьюринга. (Отметим, в частности, что действие машины Тьюринга, занятой поисками нечетного числа, являющегося суммой двух четных чисел, строго говоря, практически реализовать невозможно, так как ее детали изнашиваются гораздо раньше, чем минет вечность.) Описание любого единичного вычисления (или действия машины Тьюринга) – задача вполне конечная, а вопрос о том, завершится ли в конечном итоге это вычисление, можно полагать вполне определенным. Сначала мы доводим до логического завершения теоретические рассуждения, связанные с теми или иными идеализированными вычислениями, и лишь затем пытаемся разглядеть, каким образом наши рассуждения применимы к конечным физическим системам – таким, как реально существующие компьютеры или люди.

Ограничения конечного характера могут быть обусловлены либо тем, что (i) описание конкретного рассматриваемого вычисления оказывается слишком громоздким (т.е. число  $n$  в  $C_n$  или пара чисел  $q, n$  в  $C_q(n)$  оказываются слишком велики для того, чтобы их мог описать человек или реально существующий компьютер), либо тем, что (ii) при внешней простоте описания вычисления, тем не менее, требует для своего выполнения чрезмерно много времени, в

результате чего может показаться, что оно не завершается вовсе, хотя теоретически данное вычисление должно в конечном счете завершиться. На деле же, как мы вскоре убедимся, выясняется, что из этих двух условий сколько-нибудь существенное влияние на наши рассуждения оказывает только (i), да и оно не так уж и велико. Незначительность фактора (ii), быть может, покажется вам удивительной. Существует множество относительно простых вычислений, которые в конечном счете завершаются, однако точки их завершения путем прямого вычисления не способен достичь ни один потенциально возможный компьютер. Рассмотрим, например, следующую задачу: «распечатать последовательность из  $2^{65536}$  единиц, после чего остановиться». (В §3.26 будут предложены еще несколько подобных примеров, гораздо более интересных с математической точки зрения.) Вопрос о завершаемости того или иного вычисления не следует решать путем прямого вычисления: этот метод зачастую оказывается крайне неэффективным.

Для того, чтобы выяснить, каким образом ограничения (i) или (ii) могут повлиять на наши гёделевские рассуждения, пройдемся еще раз по соответствующим частям доказательства. В соответствии с ограничением (i), вместо бесконечного ряда вычислений, мы располагаем рядом конечным:

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

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

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

Вспомним, что утверждение (H) гласит:

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

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

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

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

Следует, впрочем, рассмотреть еще предусматриваемую ограничением (i) возможность того, что упомянутое число  $k$  окажется больше  $Q$ . В случае какого-нибудь невообразимо сложного вычисления  $A$  такая ситуация вполне возможна, однако только при условии, что это  $A$  уже начинает

приближаться к верхней границе допустимой сложности (в смысле количества двоичных знаков в его описании в формате машины Тьюринга), с которой может работать наш компьютер или человек. Это обусловлено тем, что вычисление, получающее значение  $k$  из описания вычисления  $A$  (например, в формате машины Тьюринга), – вещь достаточно простая и может быть задана в явном виде (как уже было показано в комментарии к Q6).

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

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

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

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

Не помешает несколько более подробно остановиться на сравнительно незначительном увеличении сложности, сопровождающем переход от  $A$  к  $C_k(k)$ . Помимо прочего, это существенно поможет нам в нашем дальнейшем исследовании (в §§ 3.19 и 3.20). В Приложении А (с. 191) предложено явное описание вычисления  $C_k(k)$  в виде предписаний для машины Тьюринга, рассмотренных в НРК (глава 2). Согласно этим предписаниям, под обозначением  $T_m$  понимается « $m$ -я машина Тьюринга». Для большего удобства и упрощения рассуждений здесь мы также будем пользоваться этим обозначением вместо « $C_m$ », в частности, для определения степени сложности вычислительной процедуры или отдельного вычисления. В соответствии с вышесказанным, определим степень сложности  $\mu$  машины Тьюринга  $T_m$  как количество знаков в двоичном представлении числа  $m$  (см. НРК, с. 39); при этом степень сложности некоторого вычисления  $T_m(n)$  определяется как большее из двух чисел  $\mu$  и  $\nu$ , где  $\nu$  – количество двоичных знаков в представлении числа  $n$ . Рассмотрим далее приведенное в Приложении А явное предписание для составления вычисления  $C_k(k)$  на основании алгоритма  $A$ , заданного в упомянутых спецификациях машины Тьюринга. Полагая степень сложности  $A$  равной  $\alpha$ , находим, что степень сложности явного вычисления  $C_k(k)$  не превышает числа  $\alpha + 210 \log_2(\alpha + 336)$  – а это число, в свою очередь, оказывается лишь очень незначительно больше собственно  $\alpha$ , да и то только тогда, когда число  $\alpha$  очень велико.

В вышеприведенных общих рассуждениях имеется один потенциально спорный момент. В самом деле, какой смысл рассматривать вычисления, слишком сложные даже для того, чтобы просто их записать, или те, что, будучи записанными, возможно, потребуют на свое действительное выполнение промежутков времени, гораздо больший предполагаемого возраста нашей Вселенной, даже при условии, что каждый шаг такого вычисления будет производиться за самую малую долю секунды, какая еще допускает протекание каких бы то ни было физических процессов? Упомянутое выше вычисление – то, результатом которого является последовательность из  $2^{2^{65536}}$  единиц и которое завершается лишь после выполнения этой задачи, – представля-

ет собой как раз такой пример, при этом позицию математика, позволяющего себе утверждать, что данное вычисление является незавершающимся, можно охарактеризовать как крайне нетрадиционную. Однако в математике существуют и некоторые другие точки зрения, пусть и не до такой степени нетрадиционные, – но всё же решительно презирующие всяческие условности, – согласно которым известная доля здорового скептицизма в отношении вопроса об абсолютной математической истинности идеализированных математических утверждений отнюдь не помешает. На некоторые из них, безусловно, стоит хотя бы мельком взглянуть.

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

В своем гёделевском доказательстве (в частности, в утверждении (M)) я использовал аргумент следующего вида: «Допущение о ложности  $X$  приводит к противоречию; следовательно, утверждение  $X$  истинно». Под « $X$ » в данном случае следует понимать утверждение: «Вычисление  $C_k(k)$  не завершается». Это рассуждение относится к типу *reductio ad absurdum*; что же касается доказательства Гёделя в целом, то оно и в самом деле построено именно таким образом. Направление же в математике, называемое «интуиционизмом» (у истоков которого стоял голландский математик Л.Э.Я. Брауэр; см. [223] и НРК, с. [113–116](#)), отрицает возможность построения обоснованного доказательства на основе *reductio ad absurdum*. Интуиционизм возник приблизительно в 1912 году как реакция на некоторые сформировавшиеся к концу девятнадцатого – началу двадцатого века математические тенденции, суть которых сводится к следующему: математический объект можно полагать «существующим» даже в тех случаях, когда нет никакой возможности этот объект так или иначе воплотить в действительности. А надо сказать, что слишком вольное применение крайне расплывчатой концепции математического существования и впрямь приводит порой к весьма неприятным противоречиям. Самый известный пример такого противоречия связан с парадоксальным «множеством всех множеств, не являющихся членами самих себя» Бертрана Рассела. (Если множество Рассела является членом самого себя, то оно таковым не является; если же оно членом самого себя не является, то оно им, как ни странно, является! Подробнее см. [§3.4](#) и НРК, с. [101](#).) Дабы противостоять общей тенденции, в рамках которой могут считаться «существующими» весьма вольно определенные математические объекты, интуиционисты полагают необоснованным математическое рассуждение, позволяющее делать вывод о существовании того или иного математического объекта на основании одной лишь противоречивости его несуществования. Доказательство существования объекта посредством *reductio ad absurdum* не дает абсолютно никаких оснований полагать, что упомянутый объект действительно можно построить. Каким же образом запрет на применение *reductio ad absurdum* может повлиять на наше гёделевское доказательство? Вообще говоря, совсем не может, по той простой причине, что *reductio ad absurdum* мы применяем, если можно так выразиться, наоборот, то есть противоречие в нашем случае выводится из допущения, что нечто существует, а не из обратного допущения. С интуиционистской точки зрения всё выглядит совершенно законно: мы заключаем, что объект не существует, на том основании, что противоречие возникает как раз из допущения о существовании этого самого объекта. Предложенное мною гёделевское доказательство, по сути своей, является в интуиционистском смысле абсолютно приемлемым. (См. [223], с. 492.)

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

## §2.7. Некоторые более глубокие математические соображения

Для того, чтобы лучше разобраться в значении гёделевского доказательства, полезно будет вспомнить, с какой, собственно, целью оно было первоначально предпринято. На рубеже веков ученые, деятельность которых была связана с фундаментальными математическими принципами, столкнулись с весьма серьезными проблемами. В конце XIX века – в значительной степени благодаря глубоко оригинальным математическим трудам Георга Кантора (с «диагональным доказательством» которого мы уже познакомились) – математики получили в распоряжение эффективные методы доказательства некоторых наиболее фундаментальных своих результатов, основанные на свойствах бесконечных множеств. Однако с этими преимуществами оказались связаны и не менее фундаментальные трудности, проистекающие из чересчур вольного обращения с концепцией бесконечного множества. Особо отметим парадокс Рассела (на который я вкратце ссылаюсь в комментарии к Q9, см. также §3.4 – Кантор о нем также упоминает), обозначивший некоторые препятствия, подстерегающие склонных к опрометчивым умозаключениям. Тем не менее, все понимали, что если вопрос о допустимости тех или иных методов рассуждения продумать с достаточной тщательностью, то можно добиться очень и очень впечатляющих математических результатов. Проблема, по всей видимости, сводилась к отысканию способа, посредством которого можно было бы в каждом конкретном случае абсолютно точно определить, была ли соблюдена при выборе метода рассуждения «достаточная тщательность».

Одной из главных фигур движения, поставившего перед собой цель достичь этой точности, был великий математик Давид Гильберт. Движение окрестили формализмом; в соответствии с его основополагающим принципом, следовало однозначно определить все допустимые методы математического рассуждения в пределах той или иной конкретной области раз и навсегда, включая и те, что связаны с понятием бесконечного множества. Такая совокупность правил и математических утверждений называется формальной системой. После того, как определены правила формальной системы F, решение вопроса о корректности применения этих правил – количество которых непременно является конечным<sup>145</sup> – сводится к элементарной механической проверке. Разумеется, если мы хотим, чтобы любой выводимый с помощью таких правил результат мог считаться действительно истинным, нам придется присвоить им всем статус вполне допустимых и обоснованных форм математического рассуждения. Однако некоторые из рассматриваемых правил могут подразумевать какие-либо манипуляции с бесконечными множествами, и в этом случае математическая интуиция, подсказывающая нам, какие методы рассуждения допустимы, а какие нет, может оказаться и не достойной абсолютного доверия. Сомнения в этой связи как нельзя более уместны, учитывая несоответствия, возникающие при столь вольном обращении с бесконечными множествами, что допустимым становится даже парадоксальное «множество всех множеств, не являющихся членами самих себя» Бертрана Рассела. Правила системы F не должны допускать существования «множества» Рассела, но где же, в таком случае, следует провести границу? Вообще запретить применение бесконечных множеств было бы слишком строгим ограничением (обычное евклидово пространство, например, содержит бесконечное множество точек, да и множество натуральных чисел является бесконечным); кроме того, существуют же формальные системы, абсолютно в этом смысле удовлетворительные (поскольку в их рамках не допускается, к примеру, формулировать сущности, подобные «множеству» Рассела), применяя которые можно получить большую часть необходимых математических результатов. Откуда нам знать, каким из этих формальных систем можно верить, а каким нельзя?

Рассмотрим подробнее одну такую формальную систему F; для математических утверждений, которые можно получить с помощью правил системы F, введем обозначение ИСТИННЫЕ, а для утверждений, отрицания которых выводятся из того же источника (т.е.

---

<sup>145</sup> Представление некоторых формальных систем включает в себя бесконечное количество аксиом (они описываются через посредство структур, называемых «схемами аксиом»), однако, чтобы оставаться «формальной» в том смысле, какой вкладываю в это понятие я, система должна быть выразима в каком-то конечном виде, например, упомянутая система с бесконечным количеством аксиом должна порождаться конечным набором вычислительных правил. Это вполне возможно, и именно так и обстоит дело со стандартными формальными системами, которые применяются в математических доказательствах, – одной из таких систем является, например, знаменитая «формальная система Цермело–Френкеля» ZF, описывающая традиционную теорию множеств.

утверждения, обратные рассматриваемым), – обозначение ЛОЖНЫЕ.<sup>146</sup> Любое утверждение, которое можно сформулировать в рамках системы F, но которое не является в этом смысле ни истинным, ни ложным, будем полагать НЕРАЗРЕШИМЫМ. Кто-то, возможно, сочтет, что поскольку на деле может оказаться «бессмысленным» и само понятие бесконечного множества, то, по всей видимости, нельзя абсолютно осмысленно говорить ни об истинности, ни о ложности относящихся к ним утверждений. (Это мнение применимо, по крайней мере, к некоторым разновидностям бесконечных множеств, если не ко всем.) Если придерживаться такой точки зрения, то нет особой разницы, какие именно утверждения о бесконечных множествах (некоторых разновидностей) оказываются ИСТИННЫМИ, а какие – ЛОЖНЫМИ, лишь бы не вышло так, что одно утверждение получится ИСТИННЫМ и ЛОЖНЫМ одновременно, т.е. система F должна всё же быть непротиворечивой. Собственно говоря, в этом и состоит суть истинного формализма, а в отношении формальной системы F первостепенно важно знать лишь следующее: (а) является ли она непротиворечивой и (б) является ли она полной. Система F называется полной, если любое математическое утверждение, должным образом сформулированное в рамках F, всегда оказывается либо ИСТИННЫМ, либо ЛОЖНЫМ (т.е. НЕРАЗРЕШИМЫХ утверждений система F не содержит).

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

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

В дальнейшем мы будем рассматривать только те формальные системы, которые являются достаточно обширными, чтобы содержать в себе необходимые для действительной формулировки теоремы Гёделя арифметические операции (а также, в случае нужды, и операции какой угодно машины Тьюринга; см. ниже). Говоря о какой-либо формальной системе F, я обычно буду под-

<sup>146</sup> В.Э.: Здесь лучше было бы переводчикам те слова, что они пишут заглавными буквами, оставить непереведенными как TRUE и FALSE; от этого книга только выиграла бы, а слова эти знакомы большинству читателей по многим компьютерным системам и языкам.

разумевать, что она действительно достаточно обширна в этом смысле. Это допущение не отразится на наших рассуждениях сколько-нибудь существенным образом. (Тем не менее, рассматривая формальные системы в таком контексте, я, для пущей ясности, буду иногда снабжать их эпитетом «достаточно обширная» или иным подобным.)

## §2.8. Условие $\omega$ -непротиворечивости

Наиболее известная форма теоремы Гёделя гласит, что формальная система  $F$  (достаточно обширная) не может быть одновременно полной и непротиворечивой. Это не совсем та знаменитая «теорема о неполноте», которую Гёдель первоначально представил на конференции в Кенигсберге (см. §§2.1 и 2.7), а ее несколько более сильный вариант, который был позднее получен американским логиком Дж. Баркли Россером (1936). По своей сути, первоначальный вариант теоремы Гёделя оказывается эквивалентен утверждению, что система  $F$  не может быть одновременно полной и  $\omega$ -непротиворечивой. Условие же  $\omega$ -непротиворечивости несколько строже, нежели условие непротиворечивости обыкновенной. Для объяснения его смысла нам потребуется ввести некоторые новые обозначения. В систему обозначений формальной системы  $F$  необходимо включить символы некоторых логических операций. Нам, в частности, потребуется символ, выражающий отрицание («не»); можно выбрать для этого символ « $\sim$ ». Таким образом, если  $Q$  есть некое высказывание, формулируемое в рамках  $F$ , то последовательность символов  $\sim Q$  означает «не  $Q$ ». Нужен также символ, означающий «для всех [натуральных чисел]» и называемый квантор общности; он имеет вид « $\forall$ ». Если  $P(n)$  есть некое высказывание, зависящее от натурального числа  $n$  (т.е.  $P$  представляет собой так называемую пропозициональную функцию), то строка символов  $\forall n [P(n)]$  означает «для всех натуральных чисел  $n$  высказывание  $P(n)$  справедливо». Например, если высказывание  $P(n)$  имеет вид «число  $n$  можно выразить в виде суммы квадратов трех чисел», то запись  $\forall n [P(n)]$  означает «любое натуральное число является суммой квадратов трех чисел», – что, вообще говоря, ложно (хотя, если мы заменим «трех» на «четырёх», то это же утверждение станет истинным). Такие символы можно записывать в самых различных сочетаниях; в частности, строка символов

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

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

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

$$P(0), P(1), P(2), P(3), P(4), \dots$$

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

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

В большей части предлагаемых рассуждений я не стану проводить четкую границу между непротиворечивостью и  $\omega$ -непротиворечивостью, однако тот вариант теоремы Гёделя, что

представлен в §2.5, по сути, гласит, что если формальная система  $F$  непротиворечива, то она не может быть полной, так как не может включать в себя в качестве теоремы утверждение  $G(F)$ . Здесь я всего этого демонстрировать не буду (интересующиеся же могут обратиться к [223]). Вообще говоря, для того чтобы эту форму гёделевского доказательства можно было свести к доказательству в моей формулировке, система  $F$  должна содержать в себе нечто большее, нежели просто «арифметику и обыкновенную логику». Необходимо, чтобы система  $F$  была обширной настолько, чтобы включать в себя действия любой машины Тьюринга. Иначе говоря, среди утверждений, корректно формулируемых с помощью символов системы  $F$ , должны присутствовать утверждения типа: «Такая-то машина Тьюринга, оперируя над натуральным числом  $n$ , дает на выходе натуральное число  $p$ ». Более того, имеется теорема (см. [223], главы 11 и 13), согласно которой так оно само собой и получается, если, помимо обычных арифметических операций, система  $F$  содержит следующую операцию (так называемую  $\mu$ -операцию, или операцию минимизации): «найти наименьшее натуральное число, обладающее таким-то арифметическим свойством». Вспомним, что в нашем первом вычислительном примере, (A), предложенная процедура действительно позволяла отыскать наименьшее число, не являющееся суммой трех квадратов. То есть, вообще говоря, право на подобные вещи за вычислительными процедурами следует сохранить. С другой стороны, именно благодаря этой их особенности мы и сталкиваемся с вычислениями, которые принципиально не завершаются, – например, вычисление (B), где мы пытаемся отыскать наименьшее число, не являющееся суммой четырёх квадратов, а такого числа в природе не существует.

### §2.9. Формальные системы и алгоритмическое доказательство

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

Напротив, располагая некоторой заданной вычислительной процедурой  $E$ , предназначенной для установления истинности определенных математических утверждений, мы можем построить формальную систему  $E$ , которая эффективно выражает как ИСТИННЫЕ все те истины, что можно получить с помощью процедуры  $E$ . Имеется, впрочем, и небольшая оговорка: как правило, формальная система должна содержать стандартные логические операции, однако заданная процедура  $E$  может оказаться недостаточно обширной, чтобы непосредственно включить и их. Если сама заданная процедура  $E$  не содержит этих элементарных логических операций, то при построении системы  $E$  уместно будет присоединить их к  $E$  с тем, чтобы ИСТИННЫМИ положениями системы  $E$  оказались не только утверждения, получаемые непосредственно из процедуры  $E$ , но и утверждения, являющиеся элементарными логическими следствиями утверждений, получаемых непосредственно из  $E$ . При таком построении система  $E$  не будет строго эквивалентна процедуре  $E$ , но вместо этого приобретет несколько большую мощность.

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

Поставив перед собой задачу построить на основе процедуры  $E$  формальную систему  $E$ , мы можем начать с некоторой в высшей степени фундаментальной (и, со всей очевидностью, непротиворечивой) формальной системы  $L$ , в рамках которой выражаются лишь вышеупомянутые простейшие правила логического вывода, – например, с так называемого исчисления предикатов (см. [223]), которое только на это и способно, – и построить систему  $E$  посредством присоединения к системе  $L$  процедуры  $E$  в виде дополнительных аксиом и правил процедуры для

L, переводя тем самым всякое высказывание  $P$ , получаемое из процедуры  $E$ , в разряд ИСТИННЫХ. Это, впрочем, вовсе не обязательно окажется легко достижимым на практике. Если процедура  $E$  задается всего лишь в виде спецификации машины Тьюринга, то нам, возможно, придется присоединить к системе L (как часть ее алфавита и правил процедуры) все необходимые обозначения и операции машины Тьюринга, прежде чем мы сможем присоединить саму процедуру  $E$  в качестве, по сути, дополнительной аксиомы. (См. окончание §2.8; подробности в [223].)

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

Следовательно, как процедуры для установления математических истин, алгоритмы (т.е. вычислительные процессы) и формальные системы для нужд моего доказательства, в сущности, эквивалентны.<sup>147</sup> Таким образом, несмотря на то, что представленное в §2.5 доказательство было сформулировано исключительно для вычислений, оно сходит и для общих формальных систем. В том доказательстве, если помните, речь шла о совокупности всех вычислениях (действий машины Тьюринга)  $C_q(n)$ . Следовательно, для того, чтобы оно оказалось во всех отношениях применимо к формальной системе  $F$ , эта система должна быть достаточно обширной для того, чтобы включать в себя действия всех машин Тьюринга. Алгоритмическую процедуру  $A$ , предназначенную для установления факта незавершаемости некоторых вычислений, мы можем теперь добавить к правилам системы  $F$  с тем, чтобы вычисления, предположения о незавершающемся характере которых устанавливаются в рамках  $F$  как ИСТИННЫЕ, были бы тождественны всем тем вычислениям, незавершаемость которых определяется с помощью процедуры  $A$ .

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

алгоритм  $A \leftrightarrow$  правила системы  $F$ .

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

утверждение «вычисление  $C_k(k)$  не завершается»  $\leftrightarrow$

утверждение «система  $F$  непротиворечива».

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

<sup>147</sup> В.Э.: Это одно из тех мест, где видно, как Пенроуз понимает «вычислительные процедуры». Они эквивалентны формальным системам! Ну, а то, что НЕ по формальным системам устанавливаются математические истины, – это очевидно для меня и азбука для Веданской теории. Так что в этом смысле Пенроуз для меня «ломится в открытую дверь». Но «вычислительные процедуры, которые эквивалентны формальным системам» – это еще не всё, что имеется на компьютерах. На них есть и такие программы, как упоминаемая выше программа  $N$ , классифицирующая множества по количеству элементов. И если математические истины не познаются из «вычислительных процедур, эквивалентных формальным системам», то из этого еще не следует, что эти истины не черпаются в изучении программ типа  $N$ .

есть в рамках любой обоснованной формальной системы  $F$ ), то безусловно истинно следующее утверждение:

не все предположения  $P(0), P(1), P(2), P(3), P(4), \dots$  могут быть ИСТИННЫМИ, если утверждение «предположение  $P(n)$  справедливо для всех натуральных чисел  $n$ » ЛОЖНО, что в точности совпадает с условием  $\omega$ -непротиворечивости.

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

ни одно предположение  $P$  не может быть одновременно и ИСТИННЫМ, и ЛОЖНЫМ, в точности совпадающее с условием непротиворечивости. Вообще говоря, во многих случаях различия между непротиворечивостью и  $\omega$ -непротиворечивостью практически отсутствуют. Для упрощения дальнейших рассуждений этой главы я, в общем случае, не стану разделять эти два типа непротиворечивости и буду обычно говорить просто о «непротиворечивости». Суть доказательства Гёделя и Россера сводится к тому, что установление факта непротиворечивости формальной системы (достаточно обширной) превышает возможности этой самой формальной системы. Первоначальный (кенигсбергский) вариант теоремы Гёделя опирался только на  $\omega$ -непротиворечивость, однако следующий, более известный, вывод был связан уже исключительно с непротиворечивостью обыкновенной.

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

убежденность в обоснованности равносильна убежденности в непротиворечивости.

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

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

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

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

Что касается формальной системы  $F$ , описывающей общую теорию множеств, то, действительно, не всегда ясно, можно ли вообще говорить о каком-то абсолютном смысле, в котором то или иное утверждение о множествах является либо «истинным», либо «ложным», – вследствие чего под сомнение может попасть и само понятие «обоснованности» формальной системы, подобной  $F$ . В качестве поясняющего примера приведем один известный результат, полученный Гёделем (1940) и Коэном (1966). Они показали, что определенные математические утверждения (так называемые континуум-гипотеза Кантора и аксиома выбора) никак не зависят от теоретико-множественных аксиом системы Цермело–Френкеля – стандартной формальной системы, обозначаемой здесь через  $ZF$ . (Аксиома выбора гласит, что для любой совокупности непустых множеств существует еще одно множество, которое содержит ровно один элемент из

каждого множества совокупности.<sup>148</sup> Согласно же континуум-гипотезе Кантора, количество подмножеств натуральных чисел – равное количеству вещественных чисел – представляет собой вторую по величине бесконечность после множества собственно натуральных чисел.<sup>149</sup> Читателю нет нужды вникать в скрытый смысл этих утверждений прямо сейчас. Равно как нет нужды и мне углубляться в подробное изложение аксиом и правил процедуры системы ZF.) Некоторые математики убеждены в том, что система ZF охватывает все методы математического рассуждения, необходимые для обычной математики. Некоторые даже утверждают, будто приемлемым математическим доказательством можно считать только такое доказательство, какое можно, в принципе, сформулировать и доказать в рамках системы ZF. (См. комментарий к возражению Q14, где дается оценка применимости к таким субъектам гёделевского доказательства.) Иными словами, эти математики настаивают на том, что ИСТИННЫМИ, ЛОЖНЫМИ и НЕРАЗРЕШИМЫМИ в рамках системы ZF математическими утверждениями можно считать только те утверждения, истинность, ложность и неразрешимость, соответственно, которых, в принципе, устанавливается математическими средствами. Для таких людей аксиома выбора и континуум-гипотеза являются математически неразрешимыми (что, по их мнению, и доказывалось выводом Гёделя–Козна), и они наверняка будут утверждать, что истинность или ложность этих двух математических утверждений суть предметы достаточно условные. Влияют ли эти кажущиеся неопределенности в отношении абсолютного характера математической истины на выводы, которые мы сделали из доказательства Гёделя–Тьюринга? никоим образом, так как мы имеем здесь дело с классом математических проблем гораздо более ограниченной природы, нежели те, что, подобно аксиоме выбора и континуум-гипотезе, относятся к неконструктивно-бесконечным множествам. В данном случае нас занимают лишь утверждения вида

«такое-то вычисление никогда не завершается»,

причем рассматриваемые вычисления можно задать совершенно точно через действия машины Тьюринга. Такие утверждения в логике называются Π<sub>1</sub>-высказываниями (или, точнее, Π<sub>1</sub><sup>0</sup>-высказываниями). В пределах формальной системы F утверждение G(F) является Π<sub>1</sub>-высказыванием, а вот Ω (F) таковым не является (см. §2.8). По всей видимости, не существует каких-либо разумных доводов против того, что истинный/ложный характер любого Π<sub>1</sub>-высказывания есть предмет абсолютный и никак не зависит от избранного нами мнения относительно предположений, касающихся неконструктивно-бесконечных множеств – таких, например, как аксиома выбора и континуум-гипотеза. (С другой стороны, как мы вскоре убедимся, выбор метода рассуждения, принимаемого нами в качестве инструмента для получения убедительных доказательств Π<sub>1</sub>-высказываний, действительно может определяться мнением, которого мы придерживаемся в отношении неконструктивно-бесконечных множеств; см. возражение Q11.) Очевидно, если не считать крайней позиции, занимаемой отдельными

---

<sup>148</sup> Кому-то, возможно, покажется, что это совершенно «очевидно» и уж никак не может служить предметом спора среди математиков! Проблема, однако, существует, и возникает она в связи с понятием «существования» применительно к большим бесконечным множествам. (См., например, [350], [329], [266].) На примере парадокса Рассела мы уже убедились, что в таких вопросах необходимо проявлять особую осторожность. Согласно одной точке зрения, множество не считается необходимо существующим, если нет четкого правила (не обязательно вычислимого), устанавливающего, какие элементы в это множество следует включать, а какие – нет. Как раз этого правила аксиома выбора нам и не предоставляет, поскольку в ней нет правила, определяющего, какой элемент следует взять из каждого множества совокупности. (Некоторые из следствий аксиомы выбора интуитивно не понятны и почти парадоксальны. Вероятно, в этом и состоит одна из причин возникновения разногласий по данному вопросу. Более того, я не совсем уверен, какой позиции придерживаюсь в этом отношении я сам!)

<sup>149</sup> В заключительной главе своей книги, написанной в 1966 году, Козн подчеркивает, что, хотя он и показал, что континуум-гипотеза является НЕРАЗРЕШИМОЙ в рамках процедур системы ZF, вопрос о том, является ли она действительно истинной, был оставлен им без внимания, – и выдвигает некоторые предположения относительно того, каким образом этот вопрос можно действительно решить! То есть Козн, со всей очевидностью, не считает, что выбор между принятием или непринятием континуум-гипотезы есть предмет абсолютно произвольный. Это расходится с нередко высказываемым относительно следствий из результатов Гёделя–Козна мнением, суть которого сводится к тому, что существуют многочисленные «альтернативные теории множеств», для математики в равной степени «справедливые». Такие замечания свидетельствуют о том, что Козн, подобно Гёделю, является подлинным платонистом, для которого вопросы математической истины ни в коем случае не произвольны, но абсолютны. Очень похожих взглядов придерживаюсь и я, см. §8.7.

интуиционистами (см. комментарий к Q9), единственное здоровое возражение по поводу абсолютного характера истинности таких утверждений может быть связано с тем обстоятельством, что некоторые принципиально завершающиеся вычисления могут потребовать для своего выполнения столь непомерно долгого времени, что на практике, вполне возможно, не завершатся и, скажем, за всё время жизни вселенной; может случиться и так, что для записи самого вычисления (пусть и конечного) потребуется так много символов, что физически невозможным окажется составить даже его описание. Впрочем, все эти вопросы были исчерпывающим образом проанализированы выше, в обсуждении возражения Q8, там же мы выяснили, что на наш основной вывод  $\zeta$  они никоим образом не влияют. Вспомним и о возражении Q9, рассмотрение которого показало, что позиция интуиционистов в этом случае также не избегает вывода  $\zeta$ .

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

**Q11. Существуют определенные  $\Pi_1$ -высказывания, которые можно доказать с помощью теории бесконечных множеств, однако не известно ни одного доказательства, которое использовало бы стандартные «конечные» методы. Не означает ли это, что даже к таким четко определенным проблемам математики, на деле, подходят субъективно? Различные математики, придерживающиеся в отношении теории множеств разных убеждений, могут применять к оценке математической истинности  $\Pi_1$ -высказываний неэквивалентные критерии.**

Этот момент может оказаться существенным в том, что касается моих собственных выводов из доказательства Гёделя(–Тьюринга), и я, возможно, уделил ему недостаточно много внимания в кратком изложении, представленном в НРК. Как ни странно, но возражение Q11, похоже, никого, кроме меня, не обеспокоило – по крайней мере, никто мне на него не указал! В НРК (с. 417, 418), как и здесь, я сформулировал доказательство Гёделя(–Тьюринга) исходя из того, что посредством разума и понимания способны установить все «математики» или «математическое сообщество». Преимущество подобной формулировки, в отличие от рассмотрения вопроса о способности какого-либо конкретного индивидуума к установлению математических истин посредством своего разума и понимания, заключается в том, что первый способ позволяет избежать некоторых возражений, которые нередко выдвигают в отношении той версии доказательства Гёделя, которую предложил Лукас (1961). Самые разные ученые<sup>150</sup> указывали, к примеру, на то, что «сам Лукас» никак не мог обладать знанием о своем собственном алгоритме. (Некоторые из них говорили то же самое и о варианте доказательства, предложенном мною,<sup>151</sup> не обратив, судя по всему, внимания на тот факт, что моя формулировка вовсе не настолько «личностна».) Именно возможность сослаться на способности к рассуждению и пониманию, присущие всем «математикам» вообще или «математическому сообществу», позволяет нам избежать необходимости считаться с предположением о том, что различные индивидуумы могут воспринимать математическую истину по-разному, каждый в соответствии с личным непознаваемым алгоритмом. Значительно сложнее смириться с тем, что результатом выполнения некоего непостижимого алгоритма может оказаться коллективное понимание математического сообщест-

<sup>150</sup> См., например, [202], [37].

<sup>151</sup> См., например, различные комментарии, приведенные в *Behavioral and Brain Sciences*, 13 (1990), 643–705.

ва в целом, нежели с тем, что этот самый алгоритм обуславливает математическое понимание всего лишь какого-то конкретного индивидуума. Суть возражения Q11 как раз и заключается в том, что упомянутое коллективное понимание может оказаться совсем не таким универсальным и безличным, каким счел его я.

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

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

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

Представленные мною доводы по-прежнему остаются в силе, однако, мне кажется, некоторые из более поздних утратят значительную часть своей силы, если представить ситуацию в таком виде. Более того, в случае формулировки  $\mathcal{G}^*$  всё доказательство уходит в направлении, на мой взгляд, бесперспективном, сосредоточенном, в большей степени, на конкретных механизмах, управляющих действиями конкретных индивидуумов, нежели на принципах, лежащих в основе действий любого из нас. Меня же на данном этапе интересует не столько различия подходов отдельных математиков к той или иной математической проблеме, сколько то общее, что есть между нашим пониманием и нашим математическим восприятием.

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

Тем не менее, могут быть и исключения: например, один математик полагает, что некоторое неконструктивно-бесконечное множество  $S$  «с очевидностью» существует – или, по крайней мере, что допущение о его существовании никоим образом не приводит к противоречию, – другой же математик никакой очевидности здесь не усматривает. Дискуссии математиков по таким фундаментальным вопросам могут порой принимать поистине неразрешимый характер. При этом обе стороны могут оказаться, в принципе, неспособны сколько-нибудь убедительно

изложить свои доказательства, даже в отношении  $\Pi_1$ -высказываний. Возможно, каждому математику и в самом деле присуще некое особое внутреннее восприятие истинности утверждений, связанных с неконструктивно-бесконечными множествами. Конечно же, математики нередко заявляют о том, что их восприятие таких вещей в корне отличается от восприятия коллег. Однако я полагаю, что такие различия, по сути своей, подобны различиям в ожиданиях, которые различные математики могут иметь и в отношении истинности обычных математических высказываний. Эти ожидания суть всего лишь предварительные предположения. До тех пор, пока не представлено убедительного доказательства или опровержения, математики могут спорить друг с другом об ожидаемой или предполагаемой истинности того или иного положения, однако представление такого доказательства одним из математиков убеждает (в принципе) всех. Что до фундаментальных вопросов, то там этих доказательств как раз нет. Возможно, и не будет. Быть может, их нельзя отыскать по той причине, что их просто-напросто нет, а фундаментальные вопросы допускают существование различных, но равно справедливых точек зрения. Здесь, однако, следует подчеркнуть еще один связанный с  $\Pi_1$ -высказываниями момент. Возможность наличия у математика ошибочной точки зрения – т.е. такой точки зрения, которая вынуждает его делать неверные выводы в отношении истинности тех или иных  $\Pi_1$ -высказываний, – нас в данный момент не интересует. Нет ничего невероятного в том, что математики порой опираются на неверное в фактическом отношении «понимание» – а то и на необоснованные алгоритмы, – только к настоящему обсуждению это никакого отношения не имеет, поскольку согласуется с выводом  $\zeta$ . Впрочем, эту ситуацию мы подробно рассмотрим ниже, в §3.4. Следовательно, дело в данном случае заключается не в том, могут ли разные математики придерживаться противоречащих одна другой точек зрения, а скорее в том, может ли одна точка зрения оказаться, в принципе, мощнее другой. Каждая такая точка зрения будет совершенно справедлива в том, что касается установления истинности  $\Pi_1$ -высказываний, однако какая-то из них сможет, в принципе, дать своим последователям возможность установить, что те или иные вычисления не завершаются, тогда как другие, более слабые, точки зрения на это неспособны; то есть одни математики будут обладать существенно большей способностью к пониманию, нежели другие.

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

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

а этого нам в любом случае вполне достаточно.

Есть ли другие спорные аксиомы, которые одни математики считают «очевидными», а другие ставят под сомнение? Думаю, будет огромным преувеличением сказать, что имеется хотя бы десять существенно различных точек зрения на теоретико-множественные допущения, которые в явном виде как допущения не формулируются. Положим, что их не более десяти, и

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

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

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

**Q12. Вне зависимости от того, насколько различных точек зрения придерживаются математики в принципе, на практике те же математики обладают весьма разными способностями к воспроизведению доказательств, разве не так? Не менее различны и их способности к пониманию, позволяющие им совершать математические открытия.**

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

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

Здесь я говорю исключительно о сложности математического доказательства и ни в коем случае не о возможных существенных и принципиальных вопросах, которые могут вызвать среди математиков разногласия в отношении выбора допустимых методов рассуждения. Разумеется, я

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

Чтобы еще раз подчеркнуть, что я имею в виду, скажу следующее: сам я часто посещаю математические семинары, на которых не слежу (а иногда и не пытаюсь следить) за подробностями представляемых доказательств. Наверное, если бы я сел где-нибудь и обстоятельно изучил эти самые доказательства, я и в самом деле смог бы проследить за мыслью автора – хотя, возможно, это удалось бы мне лишь при наличии дополнительной литературы или устных пояснений, которые восполнили бы возможные пробелы в моем образовании или же в материалах самого семинара. Я знаю, что в действительности я этого делать не стану. У меня почти наверняка не окажется на это ни времени, ни достаточного количества внимания, ни, впрочем, особого желания. Но при этом я вполне могу принять представленный на семинаре результат на веру по всевозможным «несущественным» причинам – например, потому что полученный результат правдоподобно «выглядит», или потому что у лектора надежная репутация, или потому что другие слушатели, которых я считаю более сведущими в таких делах, нежели я сам, этот результат оспаривать не стали. Конечно, я могу ошибиться во всех своих умозаключениях, а результат вполне может оказаться ложным – либо истинным, но никоим образом не следующим из представленного доказательства. Все эти тонкости никак не влияют на ту принципиальную позицию, которую я здесь представляю. Результат может оказаться истинным и адекватно доказанным, и в таком случае я, в принципе, могу проследить за ходом всего доказательства – или же ошибочным, в каковом случае, как уже упоминалось, он нас в данном контексте не интересует (см. §3.2 и §3.4). Возможные исключения могут составить лишь те случаи, когда представляемый материал касается каких-либо спорных аспектов теории бесконечных множеств или опирается на какой-то необычный метод рассуждения, который может быть признан сомнительным в соответствии с теми или иными математическими воззрениями (что, само по себе, может заинтриговать меня до такой степени, что я впоследствии действительно попытаюсь это доказательство повторить). Как раз такие исключительные ситуации мы обсуждали выше, в комментарии к возражению Q11. Что касается подобных соображений относительно природы математической точки зрения, на практике многие математики могут и не иметь четкого представления о том, каких именно фундаментальных принципов они в действительности придерживаются. Однако, как уже было сказано выше, в комментарии к Q11, если математик, у которого нет определенной позиции в отношении того, следует ли принимать, скажем, некую «аксиому  $Q$ », желает проявить осмотрительность, то ничто не мешает ему изложить требующие принятия аксиомы  $Q$  результаты в следующем виде: «Из принятия аксиомы  $Q$  следует, что...». Разумеется, математики, несмотря на всю их пресловутую педантичность, проявляют в подобных вопросах должную осмотрительность далеко не всегда. Нельзя отрицать и того, что время от времени им удается допускать и вовсе очевидные ошибки. И всё же все эти ошибки – если они допущены по недосмотру, а не следуют из тех или иных непоколебимых принципов – являются исправимыми. (Как упоминалось ранее, возможность действительного применения математиками в качестве основы для своих решений необоснованного алгоритма будет подробно рассмотрена в §3.2 и §3.4. Поскольку эта возможность не противоречит выводу  $\mathcal{G}$ , она не является предметом настоящего обсуждения.) В данном случае нас не занимают исправимые ошибки, так как к вопросу о принципиальной достижимости тех или иных результатов они никакого отношения не имеют. А вот возможные неопределенности в действительных взглядах математиков, безусловно, требуют дальнейшего обсуждения, которое и приводится ниже.

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

И правда, нечасто встретишь математика, способного похвалиться прочно устоявшимися и непоколебимо непротиворечивыми убеждениями, когда речь заходит об основах предмета. Кроме того, по мере накопления опыта математик вполне может изменить свои взгляды относительно того, что считать неопровержимо истинным, если он вообще склонен считать неопровержимо истинным что бы то ни было. Можно ли, например, быть совершенно и полностью уверенным в том, что число 1 отлично от числа 2? Если говорить о некоей абсолютной человеческой уверенности, то не совсем ясно, можно ли подобное понятие как-то однозначно определить. Однако какую-то точку опоры всё же выбрать необходимо. Вполне приемлемой точкой опоры может стать принятие в качестве неопровержимо истинной некоторой системы убеждений и принципов, от которой уже можно двигаться в своих рассуждениях дальше. Разумеется, нельзя забывать и о том, что многие математики вовсе не имеют определенного мнения относительно того, что именно можно считать неопровержимо истинным. Таких математиков я попросил бы какую-никакую опору для себя всё же выбрать и просто быть готовыми при необходимости впоследствии ее сменить. Как показывает доказательство Гёделя, какую бы позицию математик в этом случае ни занял, ее всё равно невозможно полностью уместить в рамки правил любой постижимой формальной системы (а если и возможно, то этот факт невозможно однозначно установить). И дело даже не в том, что та или иная конкретная позиция постоянно изменяется; система убеждений, полностью охватываемая рамками любой (достаточно обширной) формальной системы  $F$ , неизбежно должна также простирается и за пределы доступной  $F$  области. Любая позиция, среди неопровержимых убеждений которой имеется и убеждение в обоснованности системы  $F$ , должна также включать в себя и убежденность в истинности гёделевского предположения<sup>152</sup>  $G(F)$ . Убежденность в истинности  $G(F)$  не представляет собой изменения позиции; эта убежденность уже подразумевается неявно в исходной позиции, допускающей принятие истинности формальной системы  $F$ , пусть даже поначалу это и не очевидно.

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

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

Однако следует ли считать высказывание  $G(F)$  неопровержимо истинным всякий раз, когда мы признаем неопровержимо обоснованной формальную систему  $F$ ? Полагаю, утвердительный

<sup>152</sup> Пояснение к используемым здесь обозначениям можно найти в §2.8. Впрочем,  $G(F)$  без ущерба для смысла рассуждения можно было бы везде заменить на  $\Omega(F)$ , в чем мы убедимся ниже.

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

Надеюсь, читатель понимает, что возможность возникновения таких ошибок не существенна, когда речь заходит о том, что мы подразумеваем здесь под принятием предположения  $G(F)$  в качестве неопровержимой истины. Я, разумеется, говорю о действительном предположении  $G(F)$ , а не о возможном случайном предположении, непреднамеренно сформулированном благодаря опечатке или незначительной ошибке. В этой связи мне вспоминается одна история о великом американском физике Ричарде Фейнмане. Фейнман, по-видимому, объяснял одному из студентов какое-то понятие, но оговорился. Когда студент выразил недоумение, Фейнман вспылил: «Не слушайте, что я говорю; слушайте, что я имею в виду!»<sup>153</sup>.

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

Что же касается собственно моих спецификаций, следует упомянуть еще один формальный момент. Представленный мною в §2.5 вариант доказательства Гёделя(–Тьюринга) опирается не на непротиворечивость системы  $F$ , а на обоснованность алгоритма  $A$ , и являет собой критерий для установления незавершаемости вычислений (т.е. истинности  $\Pi_1$ -высказываний). Этот вариант подходит нам ничуть не хуже любых других, поскольку известно, что из обоснованности алгоритма  $A$  следует истинность утверждения о незавершаемости вычисления  $C_k(k)$ , каковое явное утверждение (тоже  $\Pi_1$ -высказывание) мы имеем полное право использовать вместо

---

<sup>153</sup> Источник цитаты мне, к сожалению, обнаружить не удалось. Однако, как справедливо заметил Рихард Йоза, точная формулировка слов Фейнмана не имеет никакого значения, поскольку послание, которое они несут, применимо и к ним самим!

высказывания  $G(F)$ . Более того, как отмечали выше (см. §2.8), доказательство, вообще говоря, зависит не от непротиворечивости формальной системы  $F$ , а от ее  $\omega$ -непротиворечивости. Из обоснованности системы  $F$  очевидно следует ее непротиворечивость, равно как и  $\omega$ -непротиворечивость. Если допустить, что система  $F$  обоснована, то ни  $\Omega(F)$ , ни  $G(F)$  из ее правил (см. §2.8) не следуют, однако оба эти высказывания являются истинными.

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

**Q14. Нет никаких сомнений в том, что формальная система  $ZF$  – или некоторая стандартная ее модификация (обозначим ее через  $ZF^*$ ) – действительно включает в себя всё необходимое для серьезной математической деятельности. Почему бы просто не принять эту систему за основу, смириться с недоказуемостью ее непротиворечивости и продолжить свои математические изыскания?**

Полагаю, такая точка зрения весьма и весьма распространена среди практикующих математиков, особенно тех, кто не слишком углубляется в фундаментальные основы или философию своего предмета. Подобное отношение вполне естественно для людей, главной заботой которых является просто хорошее выполнение серьезной, пусть и математической, работы (хотя в действительности такие люди крайне редко выражают свои результаты в рамках строгих правил формальных систем, подобных  $ZF$ ). Согласно этой точке зрения, математика имеет дело лишь с тем, что можно доказать или опровергнуть в рамках некоей конкретной формальной системы – такой, например, как  $ZF$  (или какая-либо ее модификация  $ZF^*$ ). С высоты такой позиции математическая деятельность и в самом деле напоминает своего рода «игру». Назовем ее  $ZF$ -игрой (или  $ZF^*$ -игрой), причем играть в эту игру следует в соответствии с правилами, установленными в рамках данной системы. Такой подход характерен для формалиста, подлинный же формалист мыслит исключительно в терминах ИСТИННОГО и ЛОЖНОГО, которые не обязательно совпадают с истинным и ложным в их повседневном смысле. Если формальная система обоснованна, то всё, что является ИСТИННЫМ, и будет истинным, а всё, что ЛОЖНО, будет ложным. Однако наверняка найдутся высказывания, формализуемые в рамках данной системы, которые, будучи истинными, не являются ИСТИННЫМИ, и другие, которые, будучи ложными, не являются ЛОЖНЫМИ, иными словами, в обоих случаях эти высказывания оказываются НЕРАЗРЕШИМЫМИ. Если система  $ZF$  непротиворечива, то в  $ZF$ -игре гёделевское высказывание<sup>154</sup>  $G(ZF)$  и его отрицание  $\sim G(ZF)$  принадлежат, соответственно, к этим двум категориям. (Более того, окажись система  $ZF$  противоречивой, то и высказывание  $G(ZF)$ , и его отрицание  $\sim G(ZF)$  были бы истинными и ложными одновременно!)

$ZF$ -игра, судя по всему, представляет собой исключительно разумный подход, позволяющий реализовать большую часть того, что нас интересует в обычной математике. Однако по причинам, которые обозначены выше, я совершенно не в состоянии понять, каким же образом из нее может «произрасти» реальная точка зрения в отношении чьих бы то ни было математических убеждений. Ибо если кто-то считает, что с помощью «практикуемой» им математики он устанавливает исключительно подлинные математические истины – скажем, истинность  $\Pi_1$ -высказываний, – то он должен верить и в то, что используемая им система обоснована; а если он верит в ее обоснованность, то он должен также верить в ее непротиворечивость, то есть в то, что  $\Pi_1$ -высказывание, утверждающее истинность  $G(F)$ , действительно истинно, несмотря на то, что оно НЕРАЗРЕШИМО. Таким образом, математические убеждения человека должны включать в себя нечто, что в рамках  $ZF$ -игры невыводимо. С другой стороны, если человек не верит в обоснованность формальной системы  $ZF$ , то он не может

<sup>154</sup> Как и ранее, обозначение  $G(F)$  можно без каких бы то ни было последствий заменить на  $\Omega(F)$ . То же справедливо и для комментариев к Q15–Q20.

верить и в подлинную истинность ИСТИННЫХ результатов, полученных с помощью ZF-игры. В обоих случаях сама по себе ZF-игра не в состоянии снабдить нас удовлетворительной позицией в том, что касается математической истинности. (Это равным образом применимо к любой формальной системе ZF\*).

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

Хотя этот вопрос был достаточно исчерпывающе рассмотрен в предыдущих обсуждениях, я полагаю, что суть того рассмотрения полезно будет изложить еще раз, поскольку возражения, подобные Q15, чаще всего оказываются среди нападок на наше с Лукасом приложение теоремы Гёделя. Суть же в том, что мы вовсе не утверждаем, что высказывание  $G(F)$  непременно истинно для любой формальной системы F, мы утверждаем лишь, что высказывание  $G(F)$  настолько же достоверно, насколько достоверна любая другая истина, получаемая применением правил самой системы F. (Вообще говоря, высказывание  $G(F)$  оказывается более достоверным, нежели утверждения, получаемые действительным применением правил F, так как система F, даже будучи непротиворечивой, не обязательно будет обоснованной!) Если мы верим в истинность любого утверждения P, выводимого исключительно с помощью правил системы F, то мы должны верить и в истинность  $G(F)$ , по крайней мере, в той же степени, в какой мы верим в истинность P. Таким образом, ни одна постижимая формальная система F – или эквивалентный ей алгоритм F – не может послужить абсолютно полной основой для подлинного математического познания или формирования убеждений. Как отмечалось в комментариях к Q5 и Q6, наше доказательство построено как *reductio ad absurdum*: мы выдвигаем предположение, что система F действительно является абсолютной основой для формирования убеждений; а затем показываем, что такое предположение приводит к противоречию, т.е. является неверным.

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

«высказывание P выводимо в рамках системы F»

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

«высказывание P невыводимо в рамках системы F».

Такие утверждения имеют вид: «такое-то вычисление не завершается» (или, по сути, «будучи примененным к высказыванию P, алгоритм F не завершается»), что в точности совпадает с формой рассматриваемых нами  $\Pi_1$ -высказываний. Вопрос: какие средства мы полагаем допустимыми в процессе получения подобных утверждений? Каковы, наконец, те математические процедуры, в которые мы действительно верим и применяем при установлении математических истин? Такая система убеждений, при условии, что они достаточно разумны, никак не может быть эквивалентна всего лишь убежденности в обоснованности и непротиворечивости формальной системы, какой бы эта формальная система ни была.

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

В самом деле, конечного аксиоматического способа убедиться в том, что «числа», о которых идет речь, и есть те самые подразумеваемые натуральные числа, а не какие-то посторонние «сверхнатуральные», не существует.<sup>155</sup> Однако, в некотором смысле, в этом и состоит вся суть гёделевского рассуждения. Неважно, какую именно схему аксиом формальной системы  $F$  мы построим, пытаясь охарактеризовать натуральные числа, одних лишь правил системы  $F$  будет недостаточно, чтобы определить, является ли высказывание  $G(F)$  действительно истинным или же ложным. Полагая систему  $F$  непротиворечивой, мы знаем, что в высказывании  $G(F)$  подразумевается всё же наличие некоего истинного смысла. Это, однако, происходит лишь в том случае, если символы, составляющие в действительности формальное выражение, обозначаемое « $G(F)$ », имеют подразумеваемые значения. Если эти символы интерпретировать как-либо иначе, то полученная в результате интерпретация « $G(F)$ » вполне может оказаться ложной.

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

Дело, однако, в том, что мы-то знаем, что такое на самом деле представляют собой натуральные числа, и для нас не составит никакого труда отличить от каких-то непонятных сверхнатуральных чисел. Натуральные числа суть самые обыденные вещи, обозначаемые, как правило, символами 0, 1, 2, 3, 4, 5, 6, – с этой концепцией мы знакомимся еще в детском возрасте и легко отличим ее от надуманной концепции сверхнатурального числа (см. §1.21). Есть что-то таинственное в том, что мы, похоже, и впрямь обладаем каким-то инстинктивным пониманием действительного смысла понятия натурального числа. Всё, что мы получаем в этом смысле в детском (или уже взрослом) возрасте, сводится к сравнительно небольшому количеству описаний понятий «нуля», «единицы», «двух», «трех» и т.д. («три апельсина», «один банан» и т.п.), однако при этом, несмотря на всю неадекватность такого описания, мы как-то умудряемся постичь всю концепцию в целом. В некотором платоническом смысле натуральные числа видятся своего рода категориями, обладающими абсолютным концептуальным существованием, от нас никак не зависящим. И всё же, несмотря на «человеконезависимость» натуральных чисел, мы оказываемся способны установить интеллектуальную связь с действительной концепцией натуральных чисел, опираясь лишь на неоднозначные и, на первый взгляд, неадекватные описания. С другой стороны, не существует конечного набора аксиом, с помощью которого можно было бы провести четкую границу между множеством натуральных чисел и альтернативным ему множеством так называемых «сверхнатуральных» чисел.

Более того, такое специфическое свойство всей совокупности натуральных чисел, как их бесконечное количество, мы также можем каким-то образом воспринимать непосредственно, тогда как система, действие которой ограничено точными конечными правилами, не способна отличить данную конкретную бесконечность натуральных чисел от других возможных

<sup>155</sup> Терминология была предложена Хофштадтером в [202]. Согласно «другой» теореме Гёделя – так называемой теореме о полноте, – подобные нестандартные модели существуют всегда.

(«сверхнатуральных») вариантов. Мы же легко понимаем бесконечность, характеризующую натуральные числа, пусть и обозначаем ее просто точками «...» –

«0, 1, 2, 3, 4, 5, 6, ...»,

либо сокращением «и т. д.» –

«ноль, один, два, три и т.д.».

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

Возможно, некоторые читатели знакомы с аксиомами Пеано для арифметики натуральных чисел (об арифметике Пеано я вкратце упоминал в §2.7), и, возможно, теперь эти читатели находятся в некотором недоумении: почему же аксиомы Пеано не дают адекватного определения натуральных чисел. Согласно определению Пеано, мы начинаем ряд натуральных чисел с символа 0 и затем добавляем слева особый «оператор следования», обозначаемый S и осуществляющий простое прибавление единицы к числу, над которым совершается действие, т.е. 1 определяется как S0, 2 как S1 или SS0 и т.д. В качестве правил мы располагаем следующими утверждениями: если  $Sa=Sb$ , то  $a=b$ ; и ни при каком x число 0 нельзя записать в виде Sx (последнее утверждение служит для характеристики числа 0). Кроме того, имеется «принцип индукции», согласно которому некое свойство чисел (скажем, P) должно быть истинным в отношении всех чисел n, если оно удовлетворяет двум условиям: (i) если истинно  $P(n)$ , то для всех n истинно также и  $P(Sn)$ ; (ii)  $P(0)$  истинно. Сложности начинаются, когда дело доходит до логических операций, символы которых  $\forall$  и  $\exists$  в стандартной интерпретации означают, соответственно, «для всех натуральных чисел...» и «существует такое натуральное число..., что». В нестандартной интерпретации смысл этих символов соответствующим образом изменяется, так что они квантифицируют уже не натуральные числа, а «числа» какого-то другого типа. Хотя математические спецификации Пеано, задающие оператор следования S, действительно описывают отношение упорядочения, отличающее натуральные числа от разных прочих «сверхнатуральных» чисел, эти определения невозможно записать в терминах формальных правил, которым удовлетворяют кванторы  $\forall$  и  $\exists$ . Для того, чтобы передать смысл математических определений Пеано, необходимо перейти к так называемой «логике второго порядка», в которой кванторы типа  $\forall$  и  $\exists$  также вводятся, но только теперь они оперируют не над отдельными натуральными числами, а над множествами (бесконечными) натуральных чисел. В «логике первого порядка» арифметики Пеано кванторы оперируют над отдельными числами, и в результате получается формальная система в обычном смысле этого слова. Логика же второго порядка нам формальной системы не дает. В случае строгой формальной системы вопрос о правильности применения правил системы решается чисто механическими (т.е. алгоритмическими) способами – в сущности, именно это свойство формальных систем и послужило причиной их рассмотрения в настоящем контексте. В рамках логики второго порядка упомянутое свойство не работает.

Многие ошибочно полагают (в духе приведенных в возражении Q16 соображений), что из теоремы Гёделя следует существование множества различных арифметик, каждая из которых в равной степени обоснована. Соответственно, та частная арифметика, которую мы, возможно, по чистой случайности избрали для своих нужд, определяется просто какой-то произвольно взятой формальной системой. В действительности же теорема Гёделя показывает, что ни одна из этих формальных систем (будучи непротиворечивой) не может быть полной; поэтому (как доказывается далее) к ней можно непрерывно добавлять какие угодно новые аксиомы и получать всевозможные альтернативные непротиворечивые системы, которыми при желании можно заменить ту, в рамках которой мы работаем в настоящий момент. Эту ситуацию нередко сравнивают с той, что сложилась некогда с евклидовой геометрией. На протяжении двадцати одного века люди верили, что евклидова геометрия является единственно возможной геометрией. Но, когда в восемнадцатом<sup>156</sup> веке сразу несколько великих математиков (таких как Гаусс, Лобачевский и Бойяи) показали, что существуют в равной степени возможные альтернативы общепринятой геометрии, геометрии пришлось отступить с абсолютных позиций на произвольные. Аналогично, нередко можно услышать, будто Гёдель показал, что арифметика также

<sup>156</sup> В.Э.: В девятнадцатом. (Чья ошибка? Пенроуза? Переводчика?)

представляет собой предмет произвольного выбора, при этом один набор непротиворечивых аксиом оказывается ничуть не хуже любого другого.

Однако подобная интерпретация того, что доказал Гёдель, абсолютно неверна. Согласно Гёделю, само по себе понятие формальной системы аксиом не подходит для передачи даже самых элементарных математических понятий. Когда мы употребляем термин «арифметика» без дальнейших пояснений, мы подразумеваем обычную арифметику, которая работает с обычными натуральными числами 0, 1, 2, 3, 4, ... (и, быть может, с их отрицаниями), а вовсе не со «сверхнатуральными» числами, что бы это понятие ни означало. Мы можем, если пожелаем, исследовать свойства формальных систем, и это, конечно же, станет ценным вкладом в процесс математического познания. Однако такое предприятие несколько отличается от исследования обычных свойств обычных натуральных чисел. В некотором отношении данная ситуация весьма напоминает ту, что сложилась в последнее время с геометрией. Изучение неевклидовых геометрий интересно с математической точки зрения, да и сами геометрии имеют ряд важных областей применения (например, в физике, см. НРК, глава 5, особенно рис. [5.1](#) и [5.2](#), а также [§4.4](#)), но, когда термин «геометрия» используется в обычном языке (в отличие от жаргона математиков или физиков-теоретиков), подразумевается, как правило, обычная евклидова геометрия. Однако имеется и разница: то, что логик может назвать «евклидовой геометрией», действительно можно определить (с некоторыми оговорками)<sup>157</sup> через определенную формальную систему, тогда как обычную «арифметику», как показал Гёдель, определить таким образом нельзя.

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

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

Определенные способы представления примененного к F гёделевского доказательства в рамках формальной системы F (достаточно обширной) действительно существуют, коль скоро новый, реинтерпретированный, смысл символов системы F полагается отличным от исходного смысла символов этой системы. Однако если мы пытаемся таким образом интерпретировать систему F как процедуру, с помощью которой разум приходит к тем или иным математическим выводам, то подобный подход является не чем иным, как шулерством. Если мы намерены толковать мыслительную деятельность исключительно в рамках системы F, то ее символы не должны изменять свой смысл «на полпути». Если же мы принимаем, что мыслительная деятельность может содержать что-то помимо операций самой системы F – а именно, изменение смысла символов, – то нам необходимо знать и правила, управляющие подробным изменением. Либо эти правила окажутся неалгоритмическими, и это сыграет в пользу  $\mathcal{G}$ , либо для них найдется какая-то конкретная алгоритмическая процедура, и тогда нам следовало бы изначально включить эту процедуру в нашу систему «F» – обозначим ее через  $F^+$  – с тем, чтобы она

---

<sup>157</sup> Вообще говоря, это зависит от того, какие именно утверждения считать частью так называемой «евклидовой геометрии». Если пользоваться обычной терминологией логиков, то система «евклидовой геометрии» включает только утверждения некоторого частного вида, причем оказывается, что истинность или ложность этих утверждений можно определить с помощью алгоритмической процедуры, отсюда и утверждение, что евклидову геометрию можно описать с помощью формальной системы. Однако в других интерпретациях обычная «арифметика» тоже могла бы считаться частью «евклидовой геометрии», что допустило бы классы утверждений, которые невозможно разрешить алгоритмическим путем. То же самое произошло бы, если бы мы рассмотрели задачу о замощении плоскости полимино как составляющую евклидовой геометрии, что, казалось бы, вполне естественно. В этом смысле описать геометрию Евклида формально ничуть не проще, чем арифметику!

представляла собой полную совокупность процедур, обуславливающих наши с вами понимание и проницательность, а значит, необходимости в изменении смысла символов не возникло бы вовсе. В последнем случае вместо гёделевского высказывания  $G(F)$  из предыдущего рассуждения нам предстоит разбираться уже с высказыванием  $G(F^+)$ , так что ничего мы в результате не выигрываем.

**Q18.** Даже в такой простой системе, как арифметика Пеано, можно сформулировать теорему, интерпретация которой имеет следующий смысл: «система  $F$  обоснована» следовательно «высказывание  $G(F)$  истинно». Разве это не всё, что нам нужно от теоремы Гёделя? Значит, теперь, полагая обоснованной какую угодно формальную систему  $F$ , мы вполне можем поверить и в истинность ее гёделевского высказывания – при условии, разумеется, что мы готовы принять арифметику Пеано, разве не так?

Подобную теорему<sup>158</sup> действительно можно сформулировать в рамках арифметики Пеано. Точнее (поскольку мы не можем в пределах какой бы то ни было формальной системы должным образом выразить понятие «обоснованности» или «истинности», как это следует из знаменитой теоремы Тарского), мы, в сущности, формулируем более сильный результат:

«система  $F$  непротиворечива» следовательно «высказывание  $G(F)$  истинно»,  
либо иначе:

«система  $F$   $\omega$ -непротиворечива» следовательно «высказывание  $\Omega(F)$  истинно».

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

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

«система  $N$  обоснована» следовательно «высказывание  $G(N)$  истинно»;  
или точнее, скажем так:

«система  $N$  непротиворечива» следовательно «высказывание  $G(N)$  истинно».

Если говорить о реальном смысле этих утверждений, то из них, в сущности, следует, что высказывание  $G(N)$  также утверждается системой. А так как (что касается первого из двух вышеприведенных утверждений) истинность любого производимого системой  $N$  утверждения, во всяком случае, обусловлена допущением, что система  $N$  обоснована, то получается, что если система  $N$  утверждает нечто, явно обусловленное ее собственной обоснованностью, то она вполне может утверждать это напрямую. (Из утверждения «если мне можно верить, то  $X$  истинно» следует более простое утверждение, исходящее из того же источника: « $X$  истинно».)

<sup>158</sup> См. комментарий М. Дэвиса в [74].

Однако в действительности обоснованная формальная система  $H$  не может утверждать истинность высказывания  $G(H)$ , что является следствием ее неспособности утверждать собственную обоснованность. Более того, как мы видим, она не может включать в себя и смысл символов, которыми оперирует. Те же факты годятся и для иллюстрации второго утверждения, причем в этом случае ко всему прочему добавляется и некоторая ирония: система  $H$  не способна утверждать собственную непротиворечивость лишь в том случае, если она действительно непротиворечива, если же формальная система непротиворечивой не является, то подобные ограничения ей неведомы. Противоречивая формальная система  $H$  может утверждать (в качестве «теоремы») вообще всё, что она в состоянии сформулировать! Она вполне может, как выясняется, сформулировать и утверждение: «система  $H$  непротиворечива». Формальная система (достаточно обширная) утверждает собственную непротиворечивость тогда и только тогда, когда она противоречива!

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

Когда нам дана какая-либо конкретная формальная система  $F$ , достаточно обширная и полагаемая обоснованной, мы в состоянии понять, как добавить к ней высказывание  $G(F)$  в качестве новой аксиомы и получить тем самым новую систему  $F_1$ , которая также будет считаться обоснованной. (Для согласования обозначений в последующем изложении систему  $F$  можно также обозначить через  $F_0$ .) Теперь мы можем добавить к системе  $F_1$  высказывание  $G(F_1)$ , получив в результате новую систему  $F_2$ , также, предположительно, обоснованную. Повторив данную процедуру, т.е. добавив к системе  $F_2$  высказывание  $G(F_2)$ , получим систему  $F_3$  и т.д. Приложив еще совсем немного усилий, мы непременно сообразим, как построить еще одну формальную систему  $F_\omega$ , аксиомы которой позволят нам включить в систему в качестве дополнительных аксиом для  $F$  всё бесконечное множество высказываний  $\{G(F_0), G(F_1), G(F_2), G(F_3), \dots\}$ . Очевидно, что система  $F_\omega$  также будет обоснованной. Этот процесс можно продолжить и дальше: к системе  $F_\omega$  добавляется высказывание  $G(F_\omega)$ , в результате чего получается система  $F_{\omega+1}$ , к которой затем добавляется высказывание  $G(F_{\omega+1})$ , что дает систему  $F_{\omega+2}$ , и т.д. Далее, как и в предыдущий раз, мы можем построить формальную систему  $F_{\omega^2}$  ( $=F_{\omega+\omega}$ ), включив в нее весь бесконечный набор соответствующих аксиом, каковая система опять-таки окажется очевидно обоснованной. Добавлением к ней высказывания  $G(F_{\omega^2})$ , получим систему  $F_{\omega^2+1}$  и т.д., а потом построим новую систему  $F_{\omega^3}$  ( $=F_{\omega^2+\omega}$ ), включив в нее опять-таки бесконечное множество аксиом. Повторив всю вышеописанную процедуру, мы сможем получить формальную систему  $F_{\omega^4}$ , после следующего повтора – систему  $F_{\omega^5}$  и т.д. Еще чуть-чуть потрудиться, и мы обязательно увидим, как можно включить уже это множество новых аксиом  $\{G(F_\omega), G(F_{\omega^2}), G(F_{\omega^3}), G(F_{\omega^4}), \dots\}$  в новую формальную систему  $F_{\omega^2}$  ( $=F_{\omega\omega}$ ). Повторив всю процедуру, мы получим новую систему  $F_{\omega^2+\omega^2}$ , затем – систему  $F_{\omega^2+\omega^2+\omega^2}$  и т.д.; в конце концов, когда мы сообразим, как связать всё это вместе (разумеется, и на этот раз не без некоторого напряжения умственных способностей), наши старания приведут нас к еще более всеобъемлющей системе  $F_{\omega^3}$ , которая также должна быть обоснованной.

Читатели, которые знакомы с понятием канторовых трансфинитных ординалов, несомненно, узнают индексы, обычно используемые для обозначения таких чисел. Тем же, кто от подобных вещей далек, не стоит беспокоиться из-за незнания точного значения этих символов. Достаточно сказать, что описанную процедуру «гёделизации» можно продолжить и далее: мы получим формальные системы  $F_{\omega^4}$ ,  $F_{\omega^5}$ , ..., после чего придем к еще более обширной системе  $F_{\omega^\omega}$ , затем процесс продолжается до еще больших ординалов, например,  $\omega^{\omega^\omega}$  и т.д. – до тех пор, пока мы всё еще способны на каждом последующем этапе понять, каким образом систематизировать всё множество гёделизации, которые мы получили на данный момент. В этом и заключается основная проблема: для упомянутых нами «усилий, трудов и напряжений» требуется соответствующее понимание того, как должно систематизировать предыдущие гёделизации. Эта систематизация выполнима при условии, что достигаемый к каждому последующему моменту этап будет помечаться так называемым рекурсивным ординалом, что, в сущности, означает, что должен существовать определенный алгоритм, способный такую процедуру генерировать. Однако алгоритмической процедуры, которую можно было бы заложить заранее и которая

позволила бы выполнить описанную систематизацию для всех рекурсивных ординалов раз и навсегда, просто-напросто не существует. Нам снова придется прибегать к пониманию.

Вышеприведенная процедура была впервые предложена Аланом Тьюрингом в его докторской диссертации (а опубликована в [368])<sup>159</sup>; там же Тьюринг показал, что любое истинное  $\Pi_1$ -высказывание можно, в некотором смысле, доказать с помощью многократной гёделизации, подобной описанной нами. (См. также [117].) Впрочем, воспользоваться этим для получения механической процедуры установления истинности  $\Pi_1$ -высказываний нам не удастся по той простой причине, что механически систематизировать гёделизацию невозможно. Более того, невозможность «автоматизации» процедуры гёделизации как раз и выводится из результата Тьюринга. А в §2.5 мы уже показали, что общее установление истинности (либо ложности)  $\Pi_1$ -высказываний невозможно произвести с помощью каких бы то ни было алгоритмических процедур. Так что, в поисках систематической процедуры, не доступной тем вычислительным соображениям, которые мы рассматривали до настоящего момента, многократная гёделизация нам ничем помочь не сможет. Таким образом, для вывода  $\mathcal{G}$  возражение Q19 угрозы не представляет.

**Q20. Реальная ценность математического понимания состоит, безусловно, не в том, что благодаря ему мы способны выполнять невычислимые действия, а в том, что оно позволяет нам заменить невероятно сложные вычисления сравнительно простым пониманием. Иными словами, разве не правда, что, используя разум, мы, скорее, «срезаем углы» в смысле теории сложности, а вовсе не «выскакиваем» за пределы вычислимого?**

Я вполне готов поверить в то, что на практике интуиция математика гораздо чаще используется для «обхода» вычислительной сложности, чем невычислимости. Как-никак математики по природе своей склонны к лени, а потому зачастую стараются изыскать всяческие способы избежать вычислений (пусть даже им придется в итоге выполнить значительно более сложную мыслительную работу, нежели потребовало бы собственно вычисление). Часто случается так, что попытки заставить компьютеры бездумно штамповать теоремы даже умеренно сложных формальных систем быстро загоняют эти самые компьютеры в ловушку фактически безнадежной вычислительной сложности, тогда как математик-человек, вооруженный пониманием смысла, лежащего в основе правил такой системы, без особого труда получит в рамках этой системы множество интересных результатов.<sup>160</sup>

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

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

<sup>159</sup> См. также [231], [232] и [163].

<sup>160</sup> О некоторых проблемах, с которыми сталкивались компьютерные системы, пытавшиеся самостоятельно «делать математику», можно прочесть у Д. Фридмана [124]. Отметим, что в общем случае такие системы не слишком преуспели. Они по-прежнему остро нуждаются в помощи человека.

### ***Приложение А: Гёделизирующая машина Тьюринга в явном виде***

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

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

Машина Тьюринга имеет конечное число внутренних состояний, но производит все операции на бесконечной ленте. Эта лента представляет собой линейную последовательность «ячеек», причем каждая ячейка может быть маркированной или пустой, а общее количество отметок на ленте – величина конечная. Обозначим каждую маркированную ячейку символом **1**, а каждую пустую ячейку – **0**. В машине Тьюринга имеется также считывающее устройство, которое поочередно рассматривает отметки и, в явной зависимости от внутреннего состояния машины Тьюринга и характера рассматриваемой в данный момент отметки, определяет дальнейшие действия машины по следующим трем пунктам: (i) следует ли изменить рассматриваемую в данный момент отметку; (ii) каким будет новое внутреннее состояние машины; (iii) должно ли устройство сдвинуться по ленте на один шаг вправо (обозначим это действие через  $R$ ) или влево (обозначим через  $L$ ), или же на один шаг вправо с остановкой машины ( $STOP$ ). Когда машина, в конце концов, остановится, на ленте слева от считывающего устройства будет представлен в виде последовательности символов **0** и **1** ответ на выполненное ею вычисление. Изначально лента должна быть абсолютно чистой, за исключением отметок, описывающих исходные данные (в виде конечной строки символов **1** и **0**), над которыми машина и будет выполнять свои операции. Считывающее устройство в начале работы располагается слева от всех отметок.

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

0	↔	0
1	↔	10
2	↔	100
3	↔	1010
4	↔	1000
5	↔	10010
6	↔	10100
7	↔	101010
8	↔	10000
9	↔	100010
10	↔	100100
11	↔	1001010
12	↔	101000
13	↔	1010010
14	↔	1010100
15	↔	10101010
16	↔	100000
17	↔	1000010

и т. д.

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

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

$$\begin{array}{lcl} R & \longleftrightarrow & 110 \\ L & \longleftrightarrow & 1110 \\ \text{STOP} & \longleftrightarrow & 11110. \end{array}$$

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

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

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

Согласно вышеприведенным предписаниям, в кодированной спецификации машины Тьюринга на ленте пара символов **00** должна быть представлена последовательностью **00**,

однако можно поступить более экономно и записать просто  $0$ , что явится ничуть не менее однозначным разделителем двух последовательностей, составленных из более чем одного символа  $1$  подряд.<sup>161</sup> Машина Тьюринга начинает работу, находясь во внутреннем состоянии  $0$ ; считывающее устройство движется по ленте, сохраняя это внутреннее состояние до тех пор, пока не встретит первый символ  $1$ . Это обусловлено допущением, что в набор команд машины Тьюринга всегда входит операция  $00 \rightarrow 00R$ . Таким образом, в действительной спецификации машины Тьюринга в виде последовательности  $0$  и  $1$  явного задания этой команды не требуется; вместо этого мы начнем с команды  $01 \rightarrow X$ , где  $X$  обозначает первую нетривиальную операцию запущенной машины, т.е. первый символ  $1$ , встретившийся ей на ленте. Это значит, что начальную последовательность  $110$  (команду  $\rightarrow 00R$ ), которая в противном случае непременно присутствовала бы в определяющей машину Тьюринга последовательности, можно спокойно удалить. Более того, в такой спецификации мы будем всегда удалять и завершающую последовательность  $110$ , так как она одинакова для всех машин Тьюринга.

Получаемая в результате последовательность символов  $0$  и  $1$  представляет собой самую обыкновенную (т.е. нерасширенную) двоичную запись номера машины Тьюринга  $n$  для данной машины (см. главу 2 НРК). Мы называем ее  $n$ -й машиной Тьюринга и обозначаем  $T = T_n$ . Каждый такой двоичный номер (с добавлением в конце последовательности  $110$ ) есть последовательность символов  $0$  и  $1$ , в которой нигде не встречается более четырех  $1$  подряд. Номер  $n$ , не удовлетворяющий данному условию, определяет «фиктивную машину Тьюринга», которая прекратит работать, как только встретит «команду», содержащую более четырех  $1$ . Такую машину « $T_n$ » мы будем называть некорректно определенной. Ее работа с какой угодно лентой является по определению незавершающейся. Аналогично, если действующая машина Тьюринга встретит команду перехода в состояние, определенное числом, большим всех тех чисел, для которых были явно заданы возможные последующие действия, то она также «зависнет»: такую машину мы будем полагать «фиктивной», а ее работу – незавершающейся. (Всех этих неудобств можно без особого труда избежать с помощью тех или иных технических средств, однако реальной необходимости в этом нет; СМ. §2.6, Q4).

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

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

<sup>161</sup> Это означает, что при кодировании машины Тьюринга каждую последовательность ... $110011$ ... можно заменить на ... $11011$ ... в спецификации универсальной машины Тьюринга, описанной в НРК (см. примечание 7 после главы 2), имеется пятнадцать мест, где я этого не сделал. Чрезвычайно досадная оплошность с моей стороны, и это после того, как я приложил столько усилий, чтобы добиться (в рамках моих же собственных правил) по возможности наименьшего номера, определяющего эту универсальную машину. Упомянутая простая замена позволяет уменьшить мой номер более чем в 30000 раз! Я благодарен Стивену Ганхаусу за то, что он указал мне на этот недосмотр, а также за то, что он самостоятельно проверил всю представленную в НРК спецификацию и подтвердил, что она действительно определяет универсальную машину Тьюринга.

воспользуюсь соответствующими последовательностями символов на ленте 0, 10, 110, 1110 и 11110. Таким образом, мы будем записывать

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

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

...00111110 $p$ 111110 $q$ 11111000...,

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

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

...00111110 $n$ 11111000...

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

...00111110 $n$ 111110 $n$ 11111000...

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

Явную модификацию алгоритма  $A$ , дающую такое предписание  $K$ , можно произвести следующим образом. Сначала находим в определении  $A$  начальную команду  $01 \rightarrow X$  и отмечаем для себя, что это в действительности за «X». Мы подставим это выражение вместо «X» в

спецификации, представленной ниже. Один технический момент: следует, помимо прочего, положить, чтобы алгоритм  $A$  был составлен таким образом, чтобы машина, после активации команды  $01 \rightarrow X$ , никогда больше не перешла во внутреннее состояние  $0$  алгоритма  $A$ . Это требование ни в коей мере не влечет за собой каких-либо существенных ограничений на форму алгоритма.<sup>162</sup> (Ноль можно использовать только в командах-пустышках.)

Затем при определении алгоритма  $A$  необходимо установить общее число  $N$  внутренних состояний (включая и состояние  $0$ , т.е. максимальное число внутренних состояний  $A$  будет равно  $N - 1$ ). Если в определении  $A$  нет завершающей команды вида  $(N - 1)1 \rightarrow Y$ , то в конце следует добавить команду-пустышку  $(N - 1)1 \rightarrow 00R$ . Наконец, удалим из определения  $A$  команду  $01 \rightarrow X$  и добавим ее к приводимому ниже списку машинных команд, а каждый номер внутреннего состояния, фигурирующий в этом списке, увеличим на  $N$  (символом  $\emptyset$  обозначено результирующее внутреннее состояние  $0$ , а символом «X» в записи « $11 \rightarrow X$ » представлена команда, которую мы рассмотрели выше). (В частности, первые две команды из списка примут в данном случае следующий вид:

$$01 \rightarrow N1R, N0 \rightarrow (N+4)0R.)$$

$\emptyset 1 \rightarrow 01R, 00 \rightarrow 40R, 01 \rightarrow 01R, 10 \rightarrow 21R,$   
 $11 \rightarrow X, 20 \rightarrow 31R, 21 \rightarrow \emptyset 0R, 30 \rightarrow 551R,$   
 $31 \rightarrow \emptyset 0R, 40 \rightarrow 40R, 41 \rightarrow 51R, 50 \rightarrow 40R,$   
 $51 \rightarrow 61R, 60 \rightarrow 40R, 61 \rightarrow 71R, 70 \rightarrow 40R,$   
 $71 \rightarrow 81R, 80 \rightarrow 40R, 81 \rightarrow 91R, 90 \rightarrow 100R,$   
 $91 \rightarrow \emptyset 0R, 100 \rightarrow 111R, 101 \rightarrow \emptyset 0R, 110 \rightarrow 121R,$   
 $111 \rightarrow 120R, 120 \rightarrow 131R, 121 \rightarrow 130R, 130 \rightarrow 141R,$   
 $131 \rightarrow 140R, 140 \rightarrow 151R, 141 \rightarrow 10R, 150 \rightarrow 00R,$   
 $151 \rightarrow \emptyset 0R, 160 \rightarrow 170L, 161 \rightarrow 161L, 170 \rightarrow 170L,$   
 $171 \rightarrow 181L, 180 \rightarrow 170L, 181 \rightarrow 191L, 190 \rightarrow 170L,$   
 $191 \rightarrow 201L, 200 \rightarrow 170L, 201 \rightarrow 211L, 210 \rightarrow 170L,$   
 $211 \rightarrow 221L, 220 \rightarrow 220L, 221 \rightarrow 231L, 230 \rightarrow 220L,$   
 $231 \rightarrow 241L, 240 \rightarrow 220L, 241 \rightarrow 251L, 250 \rightarrow 220L,$   
 $251 \rightarrow 261L, 260 \rightarrow 220L, 261 \rightarrow 271L, 270 \rightarrow 321R,$   
 $271 \rightarrow 281L, 280 \rightarrow 330R, 281 \rightarrow 291L, 290 \rightarrow 330R,$   
 $291 \rightarrow 301L, 300 \rightarrow 330R, 301 \rightarrow 311L, 310 \rightarrow 330R,$   
 $311 \rightarrow 110R, 320 \rightarrow 340L, 321 \rightarrow 321R, 330 \rightarrow 350R,$   
 $331 \rightarrow 331R, 340 \rightarrow 360R, 341 \rightarrow 340R, 350 \rightarrow 371R,$   
 $351 \rightarrow 350R, 360 \rightarrow 360R, 361 \rightarrow 381R, 370 \rightarrow 370R,$   
 $371 \rightarrow 391R, 380 \rightarrow 360R, 381 \rightarrow 401R, 390 \rightarrow 370R,$   
  
 $391 \rightarrow 411R, 400 \rightarrow 360R, 401 \rightarrow 421R, 410 \rightarrow 370R,$   
 $411 \rightarrow 431R, 420 \rightarrow 360R, 421 \rightarrow 441R, 430 \rightarrow 370R,$   
 $431 \rightarrow 451R, 440 \rightarrow 360R, 441 \rightarrow 461R, 450 \rightarrow 370R,$   
 $451 \rightarrow 471R, 460 \rightarrow 480R, 461 \rightarrow 461R, 470 \rightarrow 490R,$   
 $471 \rightarrow 471R, 480 \rightarrow 480R, 481 \rightarrow 490R, 490 \rightarrow 481R,$   
 $491 \rightarrow 501R, 500 \rightarrow 481R, 501 \rightarrow 511R, 510 \rightarrow 481R,$   
 $511 \rightarrow 521R, 520 \rightarrow 481R, 521 \rightarrow 531R, 530 \rightarrow 541R,$   
 $531 \rightarrow 531R, 540 \rightarrow 160L, 541 \rightarrow \emptyset 0R, 550 \rightarrow 531R.$

Теперь мы готовы точно определить предельную длину предписания  $K$ , получаемого путем вышеприведенного построения, как функцию от длины алгоритма  $A$ . Сравним эту «длину» со

<sup>162</sup> Более того, сам Тьюринг первоначально предполагал вообще останавливать машину всякий раз, когда она повторно переходит во внутреннее состояние «0» из любого другого состояния. В этом случае нам не только не понадобилось бы вышеупомянутое ограничение, мы спокойно могли бы обойтись и без команды STOP. Тем самым мы достигли бы существенного упрощения, поскольку последовательность 11110 в качестве команды нам была бы уже не нужна, и ее можно было бы использовать как разделитель, что позволило бы избавиться от последовательности 111110. Это значительно сократило бы длину предписания  $K$ , и, кроме того, вместо пятеричной системы счисления мы обошлись бы четверичной.

«степенью сложности», определенной в §2.6 (в конце комментария к возражению Q8). Для некоторой конкретной машины Тьюринга  $T_m$  (например, той, что выполняет вычисление  $A$ ) эта величина равна количеству знаков в двоичном представлении числа  $m$ . Для некоторого конкретного машинного действия  $T_m(n)$  (например, выполнения предписания  $K$ ) эта величина равна количеству двоичных цифр в большем из чисел  $m$  и  $n$ . Обозначим через  $\alpha$  и  $k$  количество двоичных цифр в  $a$  и  $k'$  соответственно, где

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

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

$$\alpha \geq 6N - 6$$

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

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

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

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

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

$$\eta = k + 13 < \alpha - 2 + 210 \log_2 (\alpha + 336),$$

или проще:

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

Детали вышеприведенного рассуждения специфичны для данного конкретного предложенного еще в НРК способа кодирования машин Тьюринга, и при использовании какого-либо иного кодирования они также будут несколько иными. Основная же идея очень проста. Более того, прими мы формализм  $\lambda$ -исчисления, вся операция оказалась бы, в некотором смысле, почти тривиальной. (Достаточно обстоятельное описание  $\lambda$ -исчисления Черча можно найти в НРК, конец главы 2; см. также [52].) Предположим, например, что алгоритм  $A$  определяется некоторым  $\lambda$ -оператором  $A$ , выполняющим действие над другими операторами  $P$  и  $Q$ , что выражается в виде операции  $(AP) Q$ . Оператором  $P$  здесь представлено вычисление  $C_p$ , а оператором  $Q$  — число  $q$ . Далее, оператор  $A$  должен удовлетворять известному требованию, согласно которому для любых  $P$  и  $Q$  должно быть истинным следующее утверждение:

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

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

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

т.е.  $KY = (AY)Y$  для любого оператора  $Y$ . Затем рассмотрим  $\lambda$ -операцию

КК.

Очевидно, что эта операция не завершается, поскольку  $КК = (АК) К$ , а завершение последней операции означало бы, что операция КК не завершается по причине принятой нами природы оператора А. Более того, оператор А не способен установить этот факт, потому что операция  $(АК) К$  не завершается. Если мы полагаем, что оператор А обладает требуемым свойством, то мы также должны предположить, что операция КК не завершается.

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

$$КК = \lambda y.(yy)(\lambda x.[(Ax)x]),$$

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

Строго говоря, это не совсем законно, поскольку в выражении для оператора А может также появиться и символ «х», и с этим нам придется что-то делать. Можно усмотреть сложность и в том, что генерируемое такой процедурой незавершающееся вычисление нельзя считать операцией над натуральными числами (поскольку вторая К в записи КК «числом» не является). Вообще говоря,  $\lambda$ -исчисление не вполне подходит для работы с явными численными операциями, и зачастую бывает довольно сложно понять, каким образом ту или иную заданную алгоритмическую процедуру, применяемую к натуральным числам, можно выразить в виде операции  $\lambda$ -исчисления. По этим и подобным причинам обсуждение с привлечением машин Тьюринга имеет, как нам представляется, более непосредственное отношение к теме нашего исследования и достигает требуемого результата более наглядным путем.

(Продолжение в книге {[PENRS2](#)})

## Послесловие ко Второй главе

2010.08.21 15:26 суббота

**В.Э.:** Поскольку я согласен с пенроузовским выводом  $\mathcal{G}$ , то его разбор возражений Q1 – Q20 почти не комментировал.

Данная в Приложении А машина Тьюринга меня тоже не интересует. Во-первых, привыкший к нормальным компьютерам, я не хочу углубляться в такое орудие каменного века и, во-вторых, и так ясно, что всегда можно написать программу, которая будет что-то искать (например,  $C_k(k)$ ) и не находить его.

Больше меня интересует, что Пенроуз скажет в Главе 3, где он обещает разбирать «невыводимость математического мышления». Возможно, там нас ожидают интересные вещи – и я спешу туда.

Первоначально я планировал поместить все три главы Первой части книги Пенроуза в этот том Векордии (PENRS1). Но оказалось, что уже первые две главы заполнили почти 120 страниц, а объем PDF файла этого тома приближается к 4 мегабайтам (когда лимит на интернетовских сайтах, где книга выставляется, – 5 мегабайтов). Так что добавление сюда еще Третьей главы сделало бы этот том слишком «тяжеловесным» (почти 200 страниц формата А4), создало бы угрозу выставлению его в Интернете и трудности для дальнейшего его возможного пополнения.

Поэтому я вынужден, вопреки логической организации материала, помещать Третью главу во Второй том – и там, видимо, уже вместе с началом Второй части книги Пенроуза...

А этот том здесь закрываем.

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 представительствами Векордии в Интернете были сайты: для русских книг – <http://vecordija.blogspot.com/>; для латышских книг – <http://vekordija.blogspot.com/>.

## Оглавление

VEcordia .....	1
Извлечение R-PENRS1 .....	1
Роджер Пенроуз .....	1
ТЕНИ РАЗУМА .....	1
Предисловие в Векордии.....	2
Роджер Пенроуз. «Тени разума» .....	3
В поисках науки о сознании.....	3
Предисловие .....	4
Благодарности .....	6
Читателю.....	7
Пролог .....	8
Часть I. Почему для понимания разума необходима новая физика? .....	11
Глава 1. Сознание и вычисление .....	11
§1.1. Разум и наука.....	11
§1.2. Спасут ли роботы этот безумный мир? .....	12
§1.3. Вычисление и сознательное мышление.....	15
§1.4. Физикализм и ментализм .....	20
§1.5. Вычисление: нисходящие и восходящие процедуры .....	20
§1.6. Противоречит ли точка зрения $\mathcal{C}$ тезису Черча–Тьюринга? .....	23
§1.7. Хаос .....	24
§1.8. Аналоговые вычисления .....	26
§1.9. Невычислительные процессы .....	28
§1.10. Завтрашний день .....	35
§1.11. Обладают ли компьютеры правами и несут ли ответственность? .....	37
§1.12. «Осознание», «понимание», «сознание», «интеллект» .....	38
§1.13. Доказательство Джона Серла .....	41
§1.14. Некоторые проблемы вычислительной модели .....	42
§1.15. Свидетельствуют ли ограниченные возможности сегодняшнего ИИ в пользу $\mathcal{C}$ ?.....	45
§1.16. Доказательство на основании теоремы Гёделя .....	49
§1.17. Платонизм или мистицизм? .....	50
§1.18. Почему именно математическое понимание? .....	51
§1.19. Какое отношение имеет теорема Гёделя к «бытовым» действиям? .....	52

§1.20. Мысленная визуализация и виртуальная реальность .....	56
§1.21. Является ли невычислимым математическое воображение?.....	58
Послесловие к Первой главе .....	60
Глава 2. Гёделевское доказательство .....	63
§2.1. Теорема Гёделя и машины Тьюринга .....	63
§2.2. Вычисления .....	65
§2.3. Незавершающиеся вычисления .....	66
§2.4. Как убедиться в невозможности завершить вычисление? .....	67
§2.5. Семейства вычислений; следствие Гёделя–Тьюринга $\mathcal{G}$ .....	71
§2.6. Возможные формальные возражения против $\mathcal{G}$ .....	80
§2.7. Некоторые более глубокие математические соображения .....	88
§2.8. Условие $\omega$ -непротиворечивости .....	91
§2.9. Формальные системы и алгоритмическое доказательство .....	92
§2.10. Возможные формальные возражения против $\mathcal{G}$ (продолжение).....	94
Приложение А: Гёделизирующая машина Тьюринга в явном виде .....	111
Послесловие ко Второй главе .....	117
Оглавление .....	118