ПРОБЛЕМА СОЗНАНИЯ В ФУНДАМЕНТАЛЬНОЙ ФИЗИКЕ

ТЕХНОЛОГИЧЕСКАЯ СИНГУЛЯРНОСТЬ, ТЕОРЕМА ПЕНРОУЗА ОБ ИСКУССТВЕННОМ ИНТЕЛЛЕКТЕ И КВАНТОВАЯ ПРИРОДА СОЗНАНИЯ

А.Д. Панов

Научно-исследовательский институт ядерной физики им. Д.В. Скобельцына Московского государственного университета им. М.В. Ломоносова

В статье критически анализируются прогнозы в отношении возможности создания сильного искусственного интеллекта (ИИ) в ближайшие десятилетия. Показано, что эти прогнозы основаны на трех плохо обоснованных предположениях и одном практически полностью непонятом обстоятельстве. Плохо обоснованными предположениями являются: 1) возможность создания сильного ИИ определяется наличием компьютеров достаточной мощности; 2) вычислительная мощность мозга определяется суммарным быстродействием синаптических связей; 3) вычислительная мощность мозга вообще может оцениваться на основе аналогии «мозг – это классический компьютер». Полностью непонятым обстоятельством является аргументация, связанная с теоремой Пенроуза об искусственном интеллекте, которая запрещает создание компьютера, обладающего всеми способностями человека, на основе архитектуры классического конечного автомата. Дается критический анализ всех этих трех предположений, теоремы Пенроуза, а также анализа следствий из теоремы Пенроуза, представленного самим Роджером Пенроузом. Выводы Пенроуза оцениваются как излишне пессимистические.

Ключевые слова: искусственный интеллект, технологическая сингулярность, теорема Гёделя–Тьюринга, конечный автомат, квантовый компьютер, вычислимость.

Введение
Искусственный интеллект и технологическая сингулярность

Проблема искусственного интеллекта (ИИ) оказалась в центре внимания общества вместе с возникновением первых компьютеров в начале 50-х гг. XX в., продолжает оставаться актуальной и сейчас, и ничто не предвещает изменения этого положения в обозримом будущем. Отношение к перспективам создания ИИ было разным. С одной стороны, всегда имелось

достаточное количество оптимистов, которые считали, что создание ИИ является чисто технической задачей, решение которой будет обеспечено ростом мощности вычислительной техники. С другой стороны, многим постепенно стало понятно, что создание ИИ не является чисто технической задачей, но является чрезвычайно сложной междисциплинарной проблемой, затрагивающей в высшей степени фундаментальные проблемы бытия. К этой последней партии принадлежит и автор настоящей статьи, по мнению которого, мы не приблизились не только к решению задачи создания настоящего сильного ИИ, но даже к внятной формулировке проблемы, которую хотим решить.

В настоящее время поляризация мнений в отношении перспектив ИИ нарастает и представлена в форме существования двух, в каком-то смысле крайних, направлений. Одно из них представлено сторонниками так называемой технологической сингулярности (ТС, подробно обсуждается ниже), другое – преимущественно Роджером Пенроузом и его последователями. Существенная неприятность текущего момента состоит в том, что между сторонниками этих крайних направлений диалог практически отсутствует. В частности, можно утверждать, что аргументы Роджера Пенроуза не поняты или даже не услышаны сторонниками ТС. Это приводит к тому, что, во-первых, аргументация сторонников ТС остается поверхностной, во-вторых, выводы самого Роджера Пенроуза остаются без адекватного критического анализа. В настоящей статье мы постараемся отчасти закрыть эту брешь, подвергнув критике, с единой (авторской) точки зрения, оба крайних направления. Анализ концепции ТС будет основан в основном на книге Рэя Курцвейла [1], анализ представлений Роджера Пенроуза на его книгах [2; 3]. Забегая вперед, отметим, что точка зрения сторонников ТС, по нашему мнению, страдает чрезмерным оптимизмом, в то время как выводы Пенроуза представляются излишне пессимистическими. В ходе анализа нам придется коснуться и некоторых проблем, которые прямо не были рассмотрены ни в рамках концепции ТС, ни в рамках концепции Пенроуза. Среди таких дополнительных вопросов следует выделить: 1) связь понятия квантовой реальности с вычислениями и с представлением вычислений компьютером; 2) вопрос о природе квантовой информации; 3) вопрос о природе «мышления» индивидуальной живой клетки и связь такого индивидуального клеточного «мышления» с работой мозга и сознанием. Разброс научных дисциплин, с которыми приходится иметь дело, как видно, очень велик, что и отражает существенно междисциплинарный характер проблемы ИИ.

Начнем с определений, или, точнее, с неформального разъяснения смысла используемых терминов. Под системами искусственного интеллекта будут пониматься автономные искусственные устройства, способные выполнять интеллектуальные функции. В этом «определении» буквально каждое слово требует уточнения и разъяснения. Например, что значит «искусственный»? Устройство, созданное другим устройством, следует ли считать искусственным? Что значит «устройство»? Является ли «устройством» ис-

кусственный геном, собранный из обычных нуклеотидов? И так далее. Мы, однако, не будем пытаться входить во все эти детали, так как, повторим, разъяснение носит неформальный характер. Главной же функцией данного определения является противопоставление искусственного интеллекта инструментальным системам, которые являются лишь средствами усиления естественного человеческого интеллекта и не могут работать автономно. Типичными примерами инструментальных систем являются компьютерные системы автоматического проектирования (САПР), различные программы компьютерной графики, редакторы текстов, базы данных и т.д. Но инструментальными системами являются также и обычный листок бумаги с карандашом, который позволяет нам провести сложные вычисления, которые мы не можем выполнить в уме; или обыкновенная книга, которая позволяет использовать огромные объемы информации, которые мы не в состоянии помнить. В этом смысле человеческий интеллект давно уже не является вполне «естественным» – он является инструментальным, но, не будучи «естественным», он не является «искусственным» в том смысле, в котором ИИ понимается в данной статье. Надо также отметить, что резкую границу между искусственным и инструментальным интеллектом провести невозможно. Инструментальная компьютерная система в некоторых фазах своего функционирования может выполнять столь большие объемы интеллектуальной работы, что приобретает в это время свойства автономного ИИ. Так что точные определения искусственного или инструментального интеллекта давать, пожалуй, бесполезно. Это напоминает абсолютно бесплодную попытку точно указать место на эволюционной лестнице, где проходит точная граница между неразумной человекообразной обезьяной и разумным человеком. Различие инструментального и автономного искусственного интеллекта – скорее количественный, но не качественный вопрос.

Под сильным искусственным интеллектом в этой статье будет пониматься такой ИИ, который превосходит человека или, по крайней мере, не уступает ему по всем интеллектуальным функциям во всех отношениях. Здесь, однако, снова требуются разъяснения и уточнения. Какой именно конкретный человек имеется в виду? Типичный? Но что это такое? Все люди разные, обладают разными интеллектуальными возможностями, при этом иных людей превзойти не так уж трудно, и по всем параметрам сразу. Для того чтобы придать дефиниции определенность, можно в отношении того уровня каждой из способностей людей, которую требуется превзойти, подразумевать максимальный уровень соответствующей способности, который можно обнаружить во всем человечестве. Чисто логическим путем приходим отсюда к тому, что под сильным ИИ следует понимать такой ИИ, который превосходит совокупную интеллектуальную мощность всего человечества по всем параметрам. Автору не приходилось видеть где-нибудь столь прямолинейной формулировки понятия сильного ИИ или какого-либо эквивалентного понятия, но, фактически, сторонники ТС неявно придерживаются именно такого определения. Заметим, что используемое здесь понятие

сильного ИИ хотя и похоже на распространенное определение [4] или на оригинальное определение Джона Сёрла [5], но является более узким и более сильным, чем это обычно принимается. Такое понятие более адекватно представлению о ТС.

Предположим теперь, что сильный ИИ в описанном выше смысле действительно будет когда-нибудь создан. Тогда в принципе люди окажутся более ненужными для дальнейшего саморазвития такого ИИ. В самом деле, для чего они, если ИИ по своим интеллектуальным возможностям превосходит все, что доступно людям? Более того, сильный ИИ может начать саморазвитие со столь высокой скоростью, что люди не только окажутся лишними в этом процессе, но и принципиально не смогут за ним уследить и понять происходящее. Будущее для людей становится полностью непонятным и непредсказуемым. Эта ситуация и называется технологической сингулярностью [1]. Термин был введен Вернором Винджем в 1993 г. [6], хотя похожие идеи высказывались неоднократно и раньше (см. для обзора [1, 6]). Ожидание технологической сингулярности порождает тревожные настроения, получившие отражение в различных публикациях, как, например, в статье Билла Джоя с характерным названием «Why the future doesn't need us» [7].

Не следует путать технологическую сингулярность с эволюционными сингулярностями разных типов. Под эволюционными сингулярностями понимаются различные (довольно многочисленные) ситуации, когда некоторый эволюционный параметр в зависимости от времени меняется в так называемом режиме с обострением: попытка экстраполяции кривой в будущее приводит за конечное время к бесконечному значению. Наиболее известна демографическая сингулярность, которая еще в 60-х годах прошлого века была обнаружена рядом авторов: Х. фон Форстером и др. [8], И.С. Шкловским [9] и др. Кривая роста народонаселения Земли до примерно 1970-го года оказывается приблизительно гиперболой, уходящей в бесконечность между 2025 и 2030 годами. Другие эволюционные кривые с обострением, приводящие к сингулярностям, тоже обычно описываются гиперболами с различными показателями. В отличие от таких эволюционных сингулярностей, технологическая сингулярность прямо ни с какими бесконечностями не связана. Термин «сингулярность» в последнем случае является метафорой и означает скорее весьма критическую ситуацию, в которой может оказаться человечество, если сильный ИИ реально когда-нибудь появится.

Основной вопрос, который возникает в связи с концепцией ТС, состоит в том, когда можно ожидать ее появление и из каких соображений надо определять эту дату. В работе Р. Курцвейла [1] дается следующий ответ на этот вопрос, который разделяют и другие сторонники ТС: как только мощность коммерческих компьютеров, выраженная в операциях в секунду (в оригинале – количество операций в секунду за одну тысячу долларов), превзойдет совокупную вычислительную мощность мозга всего человечества, сильный ИИ будет создан и технологическая сингулярность станет реальностью. Нетрудно видеть, что эта идея адресует вычислительную мощность не

отдельно взятого человеческого мозга, а именно всего человечества, что коррелирует с определением сильного ИИ, которого мы придерживаемся в статье. Для того чтобы довести этот подход до реального числа, остается ответить на два частных вопроса: 1) как будет расти вычислительная мощность коммерческих компьютеров; 2) какова вычислительная мощность мозга. Совокупная вычислительная мощность всего человечества тогда определяется просто как произведение мощности отдельного мозга на число людей, а точка пересечения кривой роста мощности компьютеров и совокупного мышления человечества (если таковая обнаружится) даст момент появления технологической сингулярности.

Для прогноза мощности компьютеров Рэй Курцвейл [1] использует так называемый закон Мура, в соответствии с которым вычислительная мощность компьютеров удваивается каждые полтора-два года. Вычислительную мощность мозга Курцвейл оценивает следующим несложным образом. Количество нейронов мозга, порядка 10^{11}, умножается на количество синаптических связей одного нейрона, масштаба тысячи, и на частоту срабатывания одной синаптической связи, около сотни Гц. Получается порядка 10^{16} операций в секунду на один мозг. Если население Земли оценить как десять миллиардов, то результирующая вычислительная мощность всего человечества будет порядка 10^{26}. Эта величина и сравнивается с кривой Мура. Точка пересечения падает приблизительно на 2045 г. – это и есть прогноз даты технологической сингулярности от Рэя Курцвейла.

Насколько обоснованным является такой прогноз? Не вполне очевидно, что закон Мура сохранит свою силу в течение достаточно длительного времени, хотя пока новые точки хорошо ложатся на кривые, использованные Курцвейлом в 2005 г., когда писалась книга [1]. Однако не закон Мура является самым слабым местом в предсказании ТС. Следует обратить внимание как минимум на три плохо обоснованных предположения и одно полностью непонятое обстоятельство. Среди плохо обоснованных предположений можно назвать следующие:

1. Возможность создания сильного ИИ определяется наличием компьютеров достаточной мощности.

2. Вычислительная мощность мозга определяется совокупным быстродействием синаптических связей нейронной сети.

3. Вычислительная мощность мозга вообще может оцениваться на основе аналогии «мозг – это классический компьютер».

До сих пор полностью непонятым обстоятельством остается теорема Пенроуза об искусственном интеллекте, которая вообще запрещает реализацию всех без исключения ментальных способностей человека на базе архитектуры классического компьютера. В последующих разделах статьи мы подробно обсудим все эти слабые места в аргументации сторонников сильного ИИ и ТС, причем особое внимание уделим разъяснению смысла теоремы Пенроуза, после чего критически рассмотрим интерпретацию следствий теоремы Пенроуза, предложенную самим Роджером Пенроузом.

1. Недооценка фактора программного обеспечения в создании сильного ИИ

«За прошедшие 15 лет “разум” наших электронных вычислительных машин улучшился в миллион раз... В течение нескольких следующих десятилетий следует ожидать увеличения характеристик “разума” машин еще по крайней мере в несколько десятков тысяч раз. “Разум” таких машин по основным параметрам будет заведомо превосходить разум человека». Звучит очень современно, не правда ли? И вполне воспроизводит суть аргументации сторонников ТС. Однако написано это было Иосифом Самуиловичем Шкловским в его знаменитой книге «Вселенная, жизнь, разум» издания 1965 г. С тех пор прошло почти 40 лет, и мощность компьютеров за это время возросла вовсе не в несколько десятков тысяч раз, о чем писал Шкловский, а более чем в миллиард (!) раз (от машин серии БЭСМ–6 и Эльбрус и их зарубежных аналогов со скоростью до 10^7 флоп до самой мощной современной супер-ЭВМ Titan, имеющей скорость $1.8 \cdot 10^{16}$ флоп, флоп – количество операций с плавающей запятой в секунду). Но где же машины, «заведомо превосходящие человеческий разум по основным параметрам»? Их нет; нет даже ничего похожего. Очевидно, что понимание ситуации было в чем-то фундаментально неверным, остается оно фундаментально неверным и сейчас. Где же ошибка?

Дело в том, что для того чтобы создать сильный ИИ, мало иметь достаточно мощное компьютерное железо. Надо знать, как это сделать. Нужны соответствующие методы, программное обеспечение, понимание того, какую именно задачу надо решить для создания сильного ИИ. Но программное обеспечение гораздо более консервативно, чем аппаратное обеспечение. Ничего похожего на закон Мура здесь нет.

Что же происходит в последние десятилетия с программным обеспечением, которое могло бы иметь отношение к созданию ИИ? В принципе, имеется два принципиально разных направления, в которых можно ожидать решения задачи. Первое направление можно назвать синтетическим. Это разработка программ, где интеллектуальные функции реализуются в основном независимо от того, как они реализованы в мозге. Они реализуются способом, наиболее адекватным архитектуре современных компьютеров. Другое направление называется обратной инженерией мозга. Здесь основная надежда возлагается на то, что, просто скопировав «в железе» функциональную структуру мозга, искусственный интеллект возникнет «сам собой». Посмотрим, какие имеются достижения в каждом из направлений.

1.1. Синтетическое направление в разработке ИИ

Состояние синтетического направления характеризуют несколько ярких примеров.

Программы для проведения аналитических вычислений (системы компьютерной алгебры) являются типичным примером современных систем ис-

кусственного интеллекта. Одной из лучших таких современных систем является программа Maxima. Более того, ряд других известных систем компьютерной алгебры, такие как Mathematica, Maple имеют то же самое интеллектуальное ядро. При этом первые версии программы Maxima (тогда она называлась Macsyma) были разработаны еще в 1972 г., и основное вычислительное ядро системы с тех пор практически не менялось. Расширялся в основном набор библиотек, совместимых с этой системой. За время существования программы мощность компьютеров возросла более чем в миллиард раз, но система Maxima как была лучшей 40 лет назад, такой и осталась. Ничего существенно нового в этой области за 40 лет создано не было.

Windows-версия популярного текстового процессора Microsoft Word появилась в 1989 году, 23 года назад. Компьютеры за это время стали почти в миллион раз быстрее, но функциональность Word практически не изменилась. Словари стали чуть полнее, да пользовательский интерфейс поменялся (по личному мнению автора – не в лучшую сторону, раньше был проще, логичнее и удобнее). Почти за 25 лет прогресса можно было бы ожидать появления чего-нибудь вроде интеллектуальных роботов – секретарей-помощников, но где они?

В начале 1990-х гг. получили широкое распространение программы-переводчики, которые, однако, переводили очень плохо. Но и современные их наследники переводят ненамного лучше. Разница состоит в основном в том, что выдается не один плохой перевод, а несколько вариантов плохих переводов, да программа размещается в Интернете, а не на дискете.

В популярных изданиях или в сети нередко можно встретить заголовки вроде «Робот-учёный делает открытия без помощи человека» (см., например, [10]). Это сильное преувеличение. Во всех таких случаях речь идет о хорошо поставленной комбинаторной задаче, для решения которой не требуются творческие способности. Творческие способности аккумулирует в себе программист, который ставит задачу и определяет, как ее надо решать.

Пожалуй, действительный прогресс наблюдается в направлении, которое известно как «искусственная жизнь – адаптивное поведение (аниматы)» (см., например, [11]). Однако и здесь результаты пока довольно ограниченные, и ни один серьезный исследователь в этой области не утверждает, что в каком-либо обозримом будущем это направление может привести к созданию сильного ИИ.

Нетрудно видеть, что стремительно возрастающее совершенство компьютерной техники, выражаемое кривой Мура, расходуется пока почти исключительно на развитие пользовательского интерфейса, миниатюризацию и телекоммуникации. В начале 1990-х персональный компьютер помещался на столе и управлялся мышью и клавиатурой, теперь он помещается в дужке очков, управляется наклоном головы и постоянно подключен к Интернету. Но интеллектуальные возможности у обоих устройств качественно не различаются. Нет сомнений, что интерфейс человек-компьютер и миниатюризация будут развиваться и дальше, но это ведет лишь к созданию человеко-

машинных инструментальных систем, в которых человек является неотъемлемой частью. Создание сильного ИИ такое развитие вовсе не приближает. Прогресс в области телекоммуникаций ведет к развитию таких новых направлений, как распределенные вычисления, чрезвычайно сильно влияет на уровень интеграции общества, но и здесь не видно никакой прямой связи с перспективами создания сильного ИИ. Скорее, люди изменяются под давлением этих новых обстоятельств. Все эти направления развития программного обеспечения не затрагивают фундаментальный вопрос о том, как должен работать сильный ИИ.

Еще одной бедой является почти катастрофический рост объема (в байтах, в строках программного кода) программного обеспечения без радикального роста его интеллектуальности (достаточно сравнить первые версии Microsoft Word, которые «весили» 1–2 мегабайта с современными мультигигабайтными дистрибутивами). Это говорит о том, что даже небольшое продвижение в интеллектуальности оплачивается экспоненциальным ростом объемов кода программ.

1.2. Обратная инженерия мозга

Хотя в направлении обратной инженерии мозга ведется реальная работа, в частности в России под руководством В.Л. Дунина-Барковского в Лаборатории обратного конструирования мозга имени Дэвида Марра [12], задача кажется очень сложной и перспективы этого направления не вполне ясны. Пессимизм вызывают результаты работ по моделированию нервной системы нематоды Caenorhabditis elegans (сокращенно C. elegans) [13].

Нематода C. elegans – крохотный червячок с длиной тела всего около миллиметра, причем C. elegans представлен особями трех полов: мужскими, женскими и гермафродитами. Мужские и женские особи имеют нервную систему, состоящую примерно из тысячи нейронов, но нервная система гермафродитов состоит из 302 нейронов. При всей простоте нервной системы нематода демонстрирует сложный репертуар поведений: навигация, поиск пищи, спаривание, обучение, социальное поведение, сон. Гермафродитные особи C. elegans являются очень удобным объектом для изучения и моделирования работы нервной системы. Вся нервная система C. elegans полностью и точно картирована, каждый нейрон имеет свое имя, причем для моделирования нервной системы, состоящей всего из 302 нейронов, проблема мощности компьютера заведомо не играет никакой роли. Однако, несмотря на то что работы по моделированию нервной системы C. elegans ведутся с начала 1990-х гг., результаты крайне ограничены. Ничего похожего на полноценную работающую модель нейронной системы C. elegans до сих пор получить не удалось, имеются только некоторые ограниченные результаты в моделировании управления движением тела нематоды. О моделировании всего репертуара поведения речи пока нет и перспективы решения этой задачи достаточно туманны.

Интересно отметить, какие именно характерные трудности встали на пути решения задачи. Во-первых, оказалось, что даже самого полного картирования нейронной системы мало. Нужно знать силу синаптических связей (пороги возбуждения каждого нейрона через каждую синаптическую связь), но они остаются неизвестными и измерить их пока не удается. Во-вторых, мало моделировать нервную систему. Чтобы понять, насколько полноценно и адекватно работает модель, нужно либо создать полноценного робота C. elegans, либо полную компьютерную модель тела (и вообще всего организма) C. elegans, чтобы наблюдать результаты работы нервной системы. В работе [13] и некоторых других статьях по этой тематике исследователи пошли по пути создания компьютерной модели тела. Но, однако, и этого оказалось мало. Тело должно существовать в среде обитания, и ее тоже нужно моделировать. В [13] с помощью уравнения Навье-Стокса моделировались жидкие среды разной вязкости. Ясно, что попытка моделирования полного репертуара поведений C. elegans встретит на этом пути огромные трудности.

Мозг человека содержит порядка ста миллиардов нейронов вместо 302 нейронов C. elegans, поэтому задача обратного конструирования мозга должна быть на много порядков сложнее. При этом имеется также задача сопряжения компьютерной модели мозга либо с телом робота-андроида, представляющего полноценную модель тела человека, либо с компьютерной моделью тела, но тогда и с компьютерной моделью всей среды обитания, которой для человека является вся Вселенная. Последний вариант показывает, что моделирование процессов одного мозга неожиданно оказывается эквивалентным моделированию всей Вселенной, что вряд ли возможно. В варианте сопряжения модели мозга с телом робота-андроида придется еще решить вопрос о мотивах поведения и вообще существования для такого моделированного мозга. Мотивами поведения нормального человека является его чувственно-эмоциональная сфера и ощущение себя частью социума. Если перенести полностью такие мотивации на робота, то результатом будет просто искусственный человек, который вынужден будет жить в реальном времени, в контакте с реальными людьми. Как удастся такому искусственному человеку примириться с наличием своих сверхспособностей? Сверхбыстрое мышление будет ему только мешать, так как выведет из мира нормальных людей. Все эти странные проблемы вместе с опытом моделирования нервной системы C. elegans не обещают решения задачи обратного конструирования человеческого мозга в обозримом будущем. Нельзя, конечно, совершенно исключить, что на этом пути еще будут найдены какие-то совсем неожиданные решения. В любом случае, опыт работы с C. elegans однозначно показывает, что на пути обратного конструирования мозга проблема отнюдь не сводится к недостаточной мощности имеющихся компьютеров, как это предполагают сторонники сильного ИИ.

Таким образом, как опыт развития синтетического направления конструирования ИИ, так и опыт обратной инженерии мозга показывает, что про-

блема состоит не только, и даже не столько в том, что для создания сильного ИИ не хватает вычислительных ресурсов, сколько в том, что непонятно, как решать задачу. И даже, собственно, непонятно, какую задачу надо решать, поскольку мы не знаем, что такое чисто человеческая способность к пониманию, что такое свобода воли – то есть непонятно, что именно нужно перенести в компьютер.

2. Цитоэтология и быстродействие мозга

Как уже отмечалось, прогнозы Рэя Курцвейла в отношении даты наступления ТС основаны на оценке быстродействия мозга просто как суммарной максимальной скорости срабатывания всех синаптических связей мозга. Между тем это есть характеристика скорости только одного из видов активности, реализуемой мозгом. Процессы, отвечающие этому виду активности, известны как быстрые процессы, и этот вид активности отвечает работе мозга как нейронной сети. Убеждение, согласно которому работа мозга как нейронной сети в основном и исчерпывает все основные функции, которые реализует мозг, называется нейронной парадигмой. С точки зрения этой парадигмы, для того чтобы смоделировать работу мозга, достаточно смоделировать работу нейронной сети мозга. Однако мозг реализует еще один (как минимум) вид активности, который отвечает не за работу нейронной сети мозга, а за модификацию структуры этой нейронной сети. Наиболее очевидными процессами этого типа являются возникновение и исчезновение синаптических связей. Процессы этого сорта называются медленными, и в настоящее время неизвестно, на каком уровне и как в мозге осуществляется управление медленными процессами. Между тем медленные процессы исключительно важны. Нет особенных сомнений в том, что быстрые процессы управляют моторикой тела, вовлечены в такие важные высшие психические функции, как речь, однако медленные процессы играют ведущую роль в обучении и других высших формах мышления, когда человек создает для себя какое-то новое понимание.

Оценка Рэя Курцвейла предполагает, что с медленными процессами в мозге вообще не связано никакое быстродействие, нуль операций в секунду. Это есть гипотеза, которая, однако, может оказаться очень далекой от истины. На то, что с медленными процессами связана чрезвычайно сложная система управления, не имеющая ничего общего с нейросетевой активностью, указывают наблюдения очень сложного поведения одноклеточных существ.

Примеров такого сложного поведения известно довольно много, и один из недавно обсуждавшихся в литературе относится к слизевику – одноклеточному колониальному существу [14]. Слизевик демонстрирует разные виды «интеллектуального поведения», среди которых, в частности, можно отметить способность проходить лабиринты, оптимизировать геометрическую форму колонии для достижения определенных целей и др. Но, пожалуй, самой удивительной оказывается способность к обучению (выработке услов-

ного рефлекса). Слизевики способны медленно перемещаться (подобно амебам), и было обнаружено, что влажный воздух заставляет их двигаться быстрее, а сухой – наоборот, замедляет перемещение. Чередуя поток влажного и сухого воздуха с определенным периодом, была обнаружена интересная особенность: перед очередной подачей сухого воздуха слизевики снижали скорость. При периодической смене влажного и сухого воздуха колония запоминала последовательность этих перемен и продолжала ее помнить несколько периодов, даже если смена потоков прекращалась. Считается, что мозг для запоминания информации использует изменение силы синаптических связей нейронной сети. Но чем запоминает слизевик, если у него вовсе нет нервной системы? Все сложное поведение слизевика, особенно описанный опыт с обучением, показывают, что существуют сложные внутриклеточные механизмы обработки информации, включая возможность обучения на основе использования внутриклеточной памяти. Этот круг явлений настолько богат и своеобразен, настолько отличается от того, что изучается стандартно понимаемой цитологией и любой другой наукой, имеющей отношение к биологии, биохимии или биофизике клетки, что его должна изучать новая, практически еще не оформившаяся наука, которую можно назвать цитоэтологией. Термин был введен В. Я. Александровым в статье [15].

Биологическая эволюция устроена таким образом, что раз обретенные находки и решения не теряются, но оказываются встроенными в весь последующий эволюционный процесс. Генетический код, появившийся у бактерий, без всяких изменений используют и высшие животные; многоклеточные живые организмы есть не что иное, как сложно организованные колонии узкоспециализированных одноклеточных организмов – клеток и т.д. Это свойство эволюционного процесса связано с такими понятиями, как аддитивность и консерватизм эволюции [16. С. 27]. Поэтому следует ожидать, что механизмы внутриклеточного управления, соответствующие уровню цитоэтологии, появившись в одноклеточном мире, унаследованы и клетками высших многоклеточных организмов. В частности, они в какой-то форме должны быть представлены в нейронах. «Аргумент от эволюции» предсказывает, что индивидуальные нейроны должны проявлять сложные формы поведения, не сводящиеся только лишь к функции нейрона как порогового переключателя в нейронной сети.

Действительно, такие виды активности, оказывается, уже давно были открыты. В статье Б.Г. Режабека [17] в эксперименте над нейрорецептором растяжения речного рака было убедительно показано, что одиночный изолированный нейрон способен к обучению. Нейрорецептор растяжения рака является чрезвычайно удобным объектом для таких опытов, так как единственный нейрон этого рецептора не связан синаптическими связями ни с одним другим нейроном и проявляет свою индивидуальную активность в очень чистом виде. Примечательно не только то, что одиночный нейрон способен обучаться, но ситуация, в которой происходило обучение нейрона в опытах Б.Г. Режабека [17], имеет очень мало общего с реальными жизнен-

ными ситуациями, с которыми приходится сталкиваться нейрону. Поэтому, помимо способности к обучению, в этих опытах была продемонстрирована еще и удивительная гибкость поведения нейрона. Хотя статья [17] – далеко не единственная, где было продемонстрировано сложное поведение индивидуального нейрона[16], но пока таких работ еще очень мало, что и говорит о самом начале становления науки – цитоэтологии. Тем не менее само существование такого сложного поведения отдельных нейронов уже практически не вызывает сомнений.

Поведение нейрона при формировании новых синапсов очень напоминает движение амебообразных одноклеточных, решающих какие-то свои собственные задачи, поэтому тот тип внутриклеточного управления, который обнаруживает себя в опытах со слизевиками [14], в опытах с единичным нейроном [17], вполне может отвечать за упомянутые выше медленные процессы мозга. Возникает естественный вопрос: какая же эффективная скорость вычислений, выраженная в операциях в секунду, может отвечать этим внутринейронным процессам управления? Чтобы ответить на этот вопрос нужно, естественно, понимать, где на субклеточном уровне может помещаться механизм «внутриклеточного сознания».

Этот важный вопрос тоже должна изучать новая наука – цитоэтология, или, точнее, молекулярная цитоэтология, но пока однозначного ответа, к сожалению, нет. По этому поводу существуют разные идеи, но мы не будем стараться дать здесь их полный обзор. Остановимся вместо этого на одной из наиболее широко обсуждаемых возможностей – системе так называемых микротрубочек, из которых построена значительная часть цитоскелета клетки. Мы будем опираться в основном на книги Роджера Пенроуза [2, 3] и его же недавнюю статью [18], написанную совместно со Стюартом Хамероффом.

Цитоскелет клетки представляет собой систему тонких белковых волокон, пронизывающих всю клетку. Функции цитоскелета в клетке невероятно сложны и разнообразны и, по всей видимости, изучены далеко не полностью. Цитоскелет – динамичная, изменяющаяся структура, в функции которой входит поддержание и адаптация формы клетки к внешним воздействиям, экзо- и эндоцитоз, обеспечение движения клетки как целого, активный внутриклеточный транспорт и клеточное деление [19]. Заметим, что экзоцитоз имеет непосредственное отношение к управлению синаптическими связями нейронов. Цитоскелет сложен структурами нескольких типов, для нас основной интерес представляют так называемые микротрубочки. Часть цитоскелета состоит из пучков таких микротрубочек. Каждая микротрубочка представляет собой, действительно, полую трубку с внешним диаметром около 25 нм, стенки ее сложены ровно из 13 рядов молекул белка-тубулина, причем молекулы в стенке уложены в правильную кристаллическую структуру. Каждая молекула тубулина представляет собой димер, состоящий из

[16] Мы не будем давать обзор работ этого направления, который сам требует отдельной статьи.

двух частей, называемых α-тубулином и β-тубулином. Молекула тубулина может находиться в двух конформациях, различающихся расстоянием между α- и β-тубулином и имеющих разный дипольный электрический момент. По этой причине молекулы тубулина могут играть роль битов с двумя состояниями. Более того, соседние молекулы тубулина в микротрубочке взаимодействуют между собой своими электрическими моментами, и благодаря правильной кристаллической структуре всей системы микротрубочка чрезвычайно напоминает клеточный автомат. Клеточный автомат, в свою очередь, может быть универсальным средством вычисления и управления (имеются доказательства возможности эмуляции универсальной машины Тьюринга клеточным автоматом). Следовательно, микротрубочки вполне могут оказаться тем универсальным вычислительным устройством, которое отвечает за сложное поведение клеток или автономных одноклеточных существ. Этим возможности микротрубочек, как носителей информации, не исчерпываются [20], но мы не имеем возможности обсуждать здесь дальнейшие детали. Какова могла бы оказаться скорость вычислений, обеспечиваемая этим устройством?

В статье [18] на основе обзора нескольких исследований утверждается, что микротрубочки демонстрируют набор резонансных частот, которые можно наблюдать в спектрах поглощения и излучения электромагнитных волн, в диапазоне от нескольких кГц до примерно 10^7 Гц. При этом именно наивысшая частота 10^7 Гц больше всего похожа на основную частоту колебаний дипольного момента молекулы тубулина, входящей в состав кристаллической решетки микротрубочки[17]. В этом случае 10^7 Гц и есть скорость срабатывания одной ячейки клеточного автомата, а быстродействие всего автомата можно получить умножив эту частоту на количество ячеек в нем. Поскольку один нейрон в составе микротрубочек содержит порядка 10^8 молекул тубулина, то полное быстродействие одного нейрона в этих терминах оказывается 10^{15} операций в секунду, а быстродействие всего мозга с его сотней миллиардов нейронов оказывается 10^{26} операций в секунду. Это на десять порядков превышает оценку быстродействия мозга, предлагаемую Рэем Курцвейлом для прогноза даты ТС, и вызывает очень большие сомнения, что закон Мура будет исправно действовать вплоть до столь огромных величин.

Естественно, речь пока не идет о том, что наличие такой скорости обработки информации внутри нейрона доказано, и не доказана даже локализация процесса внутринейронной обработки информации в микротрубочках. Однако очевидно, что возможности для огромных скоростей обработки информации внутри нейрона имеются, и они должны учитываться и очень серьезно изучаться. Надо также учитывать, что помимо микротрубочек в

[17] Альтернативным объяснением частоты 10^7 Гц является основная частота системы молекул тубулина как квантовой системы – Бозе-конденсата, что означало бы, что микротрубочка является не классическим клеточным автоматом, но квантовым клеточным автоматом, см. следующий раздел.

клетке имеются и другие кандидаты на локализацию информационных процессов. Поэтому основная оценка быстродействия мозга (10^{16} операций в секунду) представляется обоснованной очень слабо.

3. Аналогия мозг-компьютер и квантовые моды работы мозга

Методика прогноза даты появления ТС недвусмысленно подразумевает, что быстродействие компьютеров и мозга вообще можно сравнивать, то есть что мозг можно рассматривать как обычный классический компьютер, быстродействие которого может быть выражено в количестве операций в секунду. Между тем, если на некотором уровне (пусть даже весьма глубоком) в мозге происходит обработка информации в квантовом режиме, подобно тому, как это делается в квантовом компьютере, то такое сопоставление полностью утрачивает смысл.

Основное возражение против такой возможности со стороны Рэя Курцвейла состоит в том, что мозг – это место, которое очень плохо подходит для существования так называемой квантовой когерентности, необходимой для проведения квантовых вычислений. Он цитирует слова Сета Ллойда (Seth Lloyd), известного специалиста по квантовой теории и квантовой информатике, из его интервью электронному журналу Nano Magazine: «Мозг – это теплое и влажное место. Это очень неудачное окружение для использования квантовой когерентности»[18].

Аргумент Р. Курцвейла основан на неявной аналогии между опытом создания первых квантовых вычислительных устройств с квантовыми вычислителями, которые могли бы существовать в мозге. Действительно, большинство существующих прототипов квантовых компьютеров требуют для своей работы экстремально низкотемпературных условий и высокой степени изоляции устройства от окружения, и эту особенность конструкции Курцвейл неявно переносит на все вообще возможные устройства этого класса. Но такая экстраполяция вовсе не является оправданной и обоснованной.

Большинство современных прототипов квантовых компьютеров используют либо манипуляцию спиновыми состояниями (спинтроника) либо сверхпроводящие элементы. Спиновые состояния, как правило, являются очень хрупкими (подвержены декогеренции), так как отделены от окружения лишь очень малой энергетической щелью (либо не отделены практически вовсе). Поэтому сохранение таких состояний, вообще говоря, требует очень низких температур и высокой изоляции от окружения. Сверхпроводящие элементы нуждаются в низкотемпературных условиях по той простой причине, что настоящая высокотемпературная сверхпроводимость (при комнатной температуре) пока неизвестна. Однако элементы квантовых вычислительных устройств могут использовать совсем другие физические

[18] The brain is a hot, wet place. It is not a very favorable environment for exploiting quantum coherence.

принципы, которые позволят обходиться без криогеники для сохранения квантовой когерентности. Более того, недавно уже было продемонстрировано двухкубитное квантовое вычислительное устройство, хоть и являющееся спинтронным, но работающее при комнатной температуре [21]. В этом случае для с борьбы с декогеренцией была использована (и впервые продемонстрирована на практике) активная квантовая коррекция квантовых операций, но и без такой активной квантовой коррекции устойчивое существование квантово-перепутанных состояний при нормальных условиях вовсе не является чем-то необычным. Приведем несколько примеров.

Тривиальный пример представляет атом гелия. Нижняя оболочка атома заселена двумя электронами, но состояние каждого из этих электронов не описывается какой-то определенной волновой функцией. Волновую функцию имеют только оба электрона вместе, и такая волновая функция является перепутанным квантовым состоянием двух электронов. Это перепутанное состояние без всяких проблем существует не только при комнатной температуре, но и приблизительно до двадцати пяти тысяч градусов, когда электроны начинают переходить в возбужденное состояние и может произойти ионизация атома.

Менее тривиальный пример представляют молекулы красителей. Этот пример был использован Ричардом Фейнманом для иллюстрации фундаментальных положений квантовой механики в его знаменитом курсе лекций [22. С. 195–196]. На рис. 1 показана структурная формула молекулы красителя фуксина, которая может находиться в двух конформациях, в которых дипольный момент молекулы ориентирован в противоположных направлениях (|1> и |2> на рис. 1). Эти две различные конформации, хоть и имеют определенные структуры в смысле расположения химических связей, но не имеют определенной энергии. Это значит, что эти состояния нестабильны, и в данном случае нестабильность связана с возможностью спонтанного перехода конформаций друг в друга. Стационарными же состояниями оказываются два других состояния, которые представляются квантовыми суперпозициями состояний

$$|I\rangle = \frac{1}{\sqrt{2}}(|1\rangle + |2\rangle);\ \ |II\rangle = \frac{1}{\sqrt{2}}(|1\rangle - |2\rangle). \qquad (1)$$

Однако, в отличие от состояний |1> и |2>, которые имеют хоть и не определенную точно, но в среднем одинаковую энергию, энергия состояний |*I*> и |*II*> отличается на некоторую величину *E*, которая соответствует энергии электромагнитного кванта оптического диапазона. На этой частоте молекула красителя чрезвычайно активно поглощает свет (это связано с большим дипольным моментом молекулы фуксина), что и является основой действия этого вещества в качестве красителя. А само явление поглощения света является доказательством того, что молекула фуксина пребывает не в одном из состояний |1> и |2> с определенной структурной формулой, а в одной из квантово-когерентных суперпозиций этих состояний |*I*> и |*II*>. Более то-

го, если внимательно посмотреть на структуру молекулы фуксина, то нетрудно обнаружить, что состояния |*I*> и |*II*> имеют структуру так называемых состояний Белла, которые иначе называются также парами Эйнштейна–Подольского–Розена (ЭПР). Именно такие состояния и их обобщения на большее число частиц играют основную роль в квантовых вычислениях, в квантовой телепортации, и служат стандартным объектом для демонстрации парадоксов квантовой механики. Если левую кольцевую структуру обозначить буквой *A*, правую – буквой *B*, то нетрудно видеть, что разные конформации отличаются числом двойных связей в этих структурах. Обозначив вариант *A* с двумя и тремя связями как |*A*2> и |*A*3> и, аналогично, для кольца *B*, то состояния |*I*> и |*II*> можно записать как

$$|I\rangle = \frac{1}{\sqrt{2}}(|A3\rangle|B2\rangle + |A2\rangle|B3\rangle); \;\; |II\rangle = \frac{1}{\sqrt{2}}(|A3\rangle|B2\rangle - |A2\rangle|B3\rangle). \qquad (2)$$

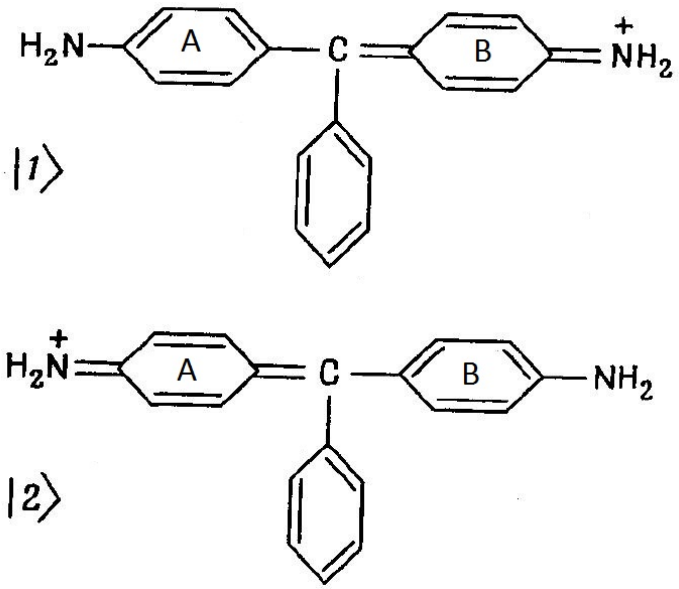

Рис. 1. Две конформации (два базисных состояния) молекулы красителя фуксина

В формуле (2) без труда узнаются скоррелированные ЭПР-пары. Для простоты мы опустили в записи состояний общий множитель, соответствующий неизменной части молекулы. Краситель фуксин прекрасно выполняет свою работу и в растворе, и в твердом состоянии, при комнатной и даже при повышенной температуре. Это означает, что в этих «неблагоприятных» условиях вполне устойчиво существуют когерентные квантово-механические суперпозиции, представляющие собой пары перепутанных состояний макромолекул.

Таким образом, в существовании квантовой спутанности даже для макромолекул при нормальных условиях, в растворе или в составе твердого тела, нет ничего необычного и, тем более, невозможного. Эволюция вполне могла бы найти способ использования таких состояний внутри живой клет-

ки, если бы только это давало какие-то селективные преимущества. Не следует недооценивать изощренность эволюции.

Можно ли указать место в нейроне, где могли бы существовать и как-то использоваться такие квантово-перепутанные состояния? Все те же упомянутые выше микротрубочки являются одним из вполне разумных кандидатов на эту роль [2; 3; 18]. Микротрубочки не просто погружены в цитоплазму клетки, но стенка микротрубочки внутри и снаружи покрыта пленкой воды, причем такой воды, которая находится в состоянии, близком к кристаллическому. Эта пленка кристаллической воды очень хорошо изолирует стенку микротрубочки, сложенную молекулами тубулина, от внешнего окружения. Поэтому нет ничего невозможного в том, что молекулы тубулина существуют не просто в одной из двух своих возможных конформаций, но в когерентной суперпозиции обеих конформаций, при этом соседние тубулины могут образовывать еще и перепутанные состояния. Тогда клеточный автомат микротрубочки будет квантовым клеточным автоматом, который может реализовать квантовый уровень управления нейрона. Есть и другие места в клетке, которые могут служить кандидатами на локализацию квантово-информационных процессов.

Таким образом, существование квантовых процессов обработки информации в мозге вовсе не исключено, хотя и не доказано прямыми наблюдениями. Однако, как мы увидим ниже, непрямое доказательство большой роли квантово-информационных процессов в работе мозга существует и связано с теоремой Пенроуза об ИИ. Если, действительно, мозг в ходе своей работы реализует что-то подобное квантовым вычислениям, то для того, чтобы превзойти мозг просто по скорости обработки информации, нужен не компьютер в обычном понимании, а некоторое устройство, в состав которого входят модули квантовых вычислений. Сколь угодно мощный классический компьютер не может симулировать и, тем более, превзойти квантовое вычислительное устройство (это будет подробнее обсуждаться ниже). Однако перспективы создания квантовых компьютеров пока не вполне ясны, хотя имеются некоторые обнадеживающие результаты: наконец, после почти тридцати лет развития идеи квантового компьютинга, создано первое полнофункциональное квантовое вычислительное устройство, поддерживающее алгоритмы квантовой коррекции кода, которые позволяют бороться с декогеренцией. Хотя устройство это состоит всего из двух кубитов [21], может быть, эти первые два кубита положат начало кривой Мура в отношении квантовых компьютеров.

В заключение этого параграфа сформулируем еще одну мысль. Как показывает проведенное выше обсуждение, нейросетевая парадигма работы мозга совсем не обязательно является истиной в последней инстанции. Это есть только одна из крайних возможностей, которые должны обсуждаться, и для полноты картины нужно сформулировать также другой предельный случай. Противоположная крайность состоит в том, что нейросетевая активность не есть носитель сознания, но есть лишь инструмент сознания, интер-

фейс, связывающий сознание с окружающей действительностью. «Само» сознание (или его жизненно важные компоненты) живет на некоторых более глубоких (субнейронных, квантовых?) уровнях организации мозга. Возможно, истина лежит где-то посередине, но ни одну из возможностей между этими двумя крайностями пока исключить нельзя.

Наконец заметим, что возможное наличие субнейронных уровней обработки информации очевидным образом резко усложняет задачу обратного конструирования мозга, в особенности если на этих уровнях присутствует квантовая обработка информации.

4. «No-go» теорема Роджера Пенроуза об искусственном интеллекте

Как мы старались показать выше, некоторые предположения, лежащие в основе прогноза появления сильного ИИ в обозримом будущем, примерно в середине XXI в., являются в лучшем случае не очень хорошо обоснованными. Но теорема Пенроуза об ИИ, которую мы обсудим ниже, предоставляет гораздо более серьезное возражение таким прогнозам.

Закон сохранения энергии (первое начало термодинамики) запрещает создание вечного двигателя первого рода. Второе начало термодинамики запрещает создание вечного двигателя второго рода. Сколь бы изощренными ни были наши технологии, устройства этого типа не могут быть созданы, так как упомянутые законы имеют характер фундаментальных запретов. Очень похоже, что роль, аналогичную первому и второму началам термодинамики в отношении вечных двигателей, относительно возможностей ИИ играет теорема Пенроуза об искусственном интеллекте.

Содержание теоремы сводится к утверждению, что какой бы мощностью ни обладало устройство, имеющее архитектуру конечного автомата (компьютера в современном понимании), человеческое мышление имеет некоторые возможности, недоступные такому устройству. Следовательно, при обсуждении возможности для ИИ превзойти человека во всех отношениях, вопрос о мощности компьютеров вообще не имеет отношения к делу до тех пор, пока мы говорим о компьютерах в современном понимании. Ни один компьютер не может превзойти мышление человека во всех отношениях независимо от его мощности, так как теорема говорит о том, что в некотором отношении человеческое мышление обязательно будет сильнее. Сосредоточимся на смысле теоремы Пенроуза, которая, как упоминалось во введении, осталась совершенно непонятой сторонниками сильного ИИ и ТС.

Начать нужно с первой теоремы Гёделя о неполноте (см., например, [23. С. 188]), на которую часто ссылаются просто как на теорему Гёделя. Смысл теоремы состоит в следующем. Пусть имеется любая непротиворечивая аксиоматическая система, содержащая в себе формальную арифметику. Тогда в этой системе существует осмысленное утверждение, которое нельзя ни доказать, ни опровергнуть средствами этой системы. Более того, доказательство теоремы имеет конструктивный характер в том смысле, что

это утверждение строится в явном виде и является истинным по построению (на метаматематическом уровне, см. [23. С. 188]). Таким образом, для любой достаточно мощной аксиоматической системы можно явно указать утверждение, про которое мы точно знаем, что оно истинно, но доказать которое в рамках этой системы невозможно. Для доказательства теоремы используются две фундаментальные идеи: так называемая Гёделевская нумерация и диагональный метод Кантора. Сама теорема Гёделя в высшей степени нетривиальна, содержит в себе множество тонкостей (некоторые детали обсуждаются в нашей статье [24]), но к настоящему времени чрезвычайно подробно исследована и не вызывает никаких сомнений. Так что она является вполне надежным исходным пунктом для рассуждений.

Структура любого конечного автомата (компьютера) может быть описана конечным образом просто в силу конечности этого устройства. Это описание аналогично конечному набору аксиом некоторой формальной системы. Предполагается также, что автомат реализует обоснованные процедуры, то есть такие процедуры, которые дают правильные ответы либо не дают ответа вовсе. Это означает, что если автомат вычисляет, например, A + B и получает C, значит, так оно и есть, этот ответ правильный. Отсутствие ответа означает, что автомат не может завершить процедуру, впадая в какую-то разновидность бесконечного цикла. Обоснованность процедур играет для конечного автомата ту же роль, что непротиворечивость для системы аксиом. Требуется также, чтобы автомат был настолько гибким и сильным, чтобы в нем можно было реализовать алгоритм, предназначенный для анализа других алгоритмов той же машины на предмет их остановки (не впадают ли в бесконечный цикл). Это требование аналогично требованию достаточной силы аксиоматической системы в теореме Гёделя. С точки зрения последующего сопоставления возможностей машины и человека в этом требовании нет ничего противоестественного. Действительно, человек может ставить перед собой такие проблемы и разрешать их. Если машина этого не может, то она заведомо слабее человека и просто не представляет интереса для анализа. Всем этим требованиям удовлетворяет универсальная машина Тьюринга и любое эквивалентное ей устройство. В частности, обычные наши вычислительные машины, которые имеют архитектуру фон Неймана, функционально эквивалентны универсальной машине Тьюринга, то есть заведомо входят в область анализа.

Как и следует ожидать, в полной аналогии с теоремой Гёделя, для такой системы (машины) можно явно построить некоторое истинное утверждение, истинность которого не может быть доказана (точнее говоря, вычислена) данным конечным автоматом. В этом заключается смысл теоремы Гёделя–Тьюринга для конечных автоматов. Как и теорема Гёделя, теорема Гёделя–Тьюринга доказывается с использованием метода, очень похожего на Гёделевскую нумерацию, и с использованием диагонального метода Кантора. То есть имеет место полная аналогия с теоремой Гёделя даже в деталях доказательства, и существование самой теоремы Гёделя–Тьюринга о конечных ав-

томатах не вызывает никакого удивления. Подобно доказательству теоремы Гёделя, доказательство теоремы Гёделя–Тьюринга тоже является конструктивным. Утверждение, истинность которого не может быть проверена конечным автоматом, строится в явном виде для любого конечного автомата и является истинным по построению. Собственно говоря, теорема Гёделя–Тьюринга и является просто теоремой Гёделя, переформулированной для конечных автоматов. Существование теоремы, аналогичной теореме Гёделя, но для конечных автоматов, а не для формальных систем, можно было предвидеть заранее, так как универсальная машина Тьюринга представляет собой, по построению, универсальное средство формализации любых вычислений, поэтому она в каком-то смысле эквивалентна полной арифметике, и все теоремы арифметики будут справедливы в отношении универсальной машины Тьюринга в надлежащей переформулировке. Таким образом, теорема Гёделя–Тьюринга, как и теорема Гёделя, является очень надежным шагом в анализе. С деталями доказательства теоремы Гёделя–Тьюринга можно ознакомиться по книгам Роджера Пенроуза [2; 3] и по специальной литературе, но эти детали не играют существенной роли в нашем изложении.

Чрезвычайно важно, что из одной только теоремы Гёделя–Тьюринга о конечных автоматах невозможно сделать прямых выводов о сопоставлении возможностей ИИ и интеллекта человека. Теорема Гёделя–Тьюринга ничего не говорит об интеллектуальных возможностях человека. Именно это обстоятельство не понято сторонниками сильного ИИ (см. обсуждение в конце этого параграфа). Здесь необходим еще один весьма нетривиальный шаг, и Роджер Пенроуз этот шаг делает. Подчеркнем еще раз, что этот последний шаг был сформулирован именно Роджером Пенроузом, он не тождествен теореме Гёделя–Тьюринга; это нечто существенно большее, при всей кажущейся простоте этого последнего шага.

Рассуждение Пенроуза представляет собой доказательство от противного. Предположим, что создан суперкомпьютер, имеющий архитектуру конечного автомата, который реализует как минимум все методы математических рассуждений, которыми владеет человечество (сильный математический ИИ). Однако, в силу теоремы Гёделя–Тьюринга, для данного суперкомпьютера, как и для любого конечного автомата, любой человек, понимающий теорему Гёделя–Тьюринга, используя математическое рассуждение, зафиксированное в этой теореме, может явно построить утверждение, про которое ему будет точно известно, что оно истинно (по построению), хотя для данного суперкомпьютера его истинность недоступна. Мы получили противоречие: предположив, что суперкомпьютер владеет всеми методами математических рассуждений, которыми владеют люди, мы немедленно указали математическое рассуждение, которое доступно любому человеку, понимающему теорему Гёделя–Тьюринга, но недоступно этому компьютеру. То есть мы доказали, что компьютер владеет не всеми методами математических рассуждений людей. Противоречие доказывает, что исходное предположение было неверным, следовательно, суперкомпьютер, владею-

щий всеми математическими способностями людей, невозможен, тем самым некоторые способности людей остаются за пределами достижимости любого вычислительного устройства – конечного автомата. Сильный ИИ невозможен ни для каких компьютеров на основе архитектуры конечного автомата.

В последнем абзаце доказательство Пенроуза приведено полностью. Краткость и простота доказательства являются обманчивыми, здесь есть несколько тонкостей, которые необходимо сразу отметить и обсудить.

Во-первых, какова природа теоремы Пенроуза? Является ли она математической теоремой в точном смысле слова? Если нет, то чем именно она является? Настораживает то, что теорема оперирует таким, например, явно не математическим понятием, как «все человечество». Каким образом «все человечество» может быть элементом математического рассуждения, если нормальные математические доказательства адресуют абстрактные математические объекты?

Если все люди смертны и Сократ – человек, то Сократ смертен. Сократ не является абстрактным математическим объектом, однако приведенное рассуждение является правильным математическим доказательством. К какой же области математики относится такое доказательство? Это рассуждение относится к математической логике, которая является разделом математики, который изучает отношения между высказываниями произвольной природы, полностью абстрагируясь от природы объектов, к которым относятся эти высказывания. Роль играет только их структура. Именно с этой точки зрения легко может быть проанализирована теорема Пенроуза. Пусть x означает «способ математических рассуждений, доступный кому-то из людей», предикат $P(x)$ означает «x может быть воспроизведен компьютером». Тогда нетрудно понять, что, в стандартных обозначениях математической логики, все доказательство Пенроуза может быть представлено короткой формулой

$$(\forall x P(x) \Rightarrow \exists y \neg P(y)) \Rightarrow \neg(\forall x P(x)) \qquad (3)$$

Эта формула представляет собой теорему математической логики (тавтологию), которая является истинной совершенно независимо от семантики, которая вкладывается в символы x, y, P. То, что в семантике возникает довольно сложное понятие «человечество», не играет роли с точки зрения истинности всего вывода. Так что на природу теоремы Пенроуза возможен взгляд как на теорему математической логики или просто как на правильный математический силлогизм.

Второй тонкий момент заключается в том, что теорема Гёделя–Тьюринга, строго говоря, адресует не реальные вычислительные машины, а абстрактные конечные автоматы – машины Тьюринга (это очень распространенный аргумент противников теоремы Гёделя–Тьюринга). Поэтому получается, что теорема Гёделя–Тьюринга как бы вовсе и не имеет отношения к реальной вычислительной технике, а имеет отношение только к каким-

то абстрактным математическим объектам. Теорема Пенроуза тогда наследует этот недостаток.

В этом рассуждении имеется методическая ошибка. Действительно, абстрактные машины Тьюринга не являются реальными вычислительными устройствами, но являются только идеальными моделями таких устройств. С другой стороны, если мы хотим оставаться в рамках науки, то надо ясно понимать, что у нас нет иного пути изучения реальности, кроме как на языке идеальных математических моделей. Если вы хотите иметь дело с реальностью напрямую, то вам не остается ничего иного, кроме молчаливого созерцания в духе буддизма. Поэтому использование идеальной модели для понимания работы реальных вычислительных машин не является чем-то необычным или неправильным. Другое дело, что кто-нибудь может сказать, что машина Тьюринга является недостаточно хорошей моделью реальных машин в контексте данной задачи. Реальные машины дают сбои, могут вовсе сломаться, обладают только конечной памятью и т.д. Но если такой недовольный захочет доказать свою правоту, то ему придется построить модель машины, которая работает со сбоями и ошибками (снова идеальную), и с использованием этой модели показать, например, что такая машина может решить задачи, недоступные машине, которая работает идеально правильно. Интересно было бы увидеть такое доказательство.

Следует отметить, что теорема Пенроуза существенно опирается на метод доказательства от противного (что, впрочем, можно сказать и о теоремах Гёделя и Гёделя–Тьюринга, что прямо видно из их доказательств). Этот метод принят в основной части нормальной математики, и огромное количество результатов находится от него в прямой зависимости. Однако не все исследователи оснований математики его принимают. Они, в частности, будут настаивать, на том, что если существование некоторого объекта приводит к противоречию, то это не означает автоматически, что объект не существует. Несуществование некоторого объекта должно быть доказано явным перечислением всех объектов того же типа, которые существуют, и отсутствием среди них искомого. Такое направление в основаниях математики известно как интуиционизм или конструктивизм (это практически синонимы). В конструктивной и интуиционистской логике формула (3) не является истинной теоремой. С точки зрения конструктивной математики, для доказательства теоремы Пенроуза нужно было бы представить список всех возможных конечных автоматов и явно указать, что ни один из них не превосходит все математические способности людей. Вряд ли такое доказательство возможно. Таким образом, теорема Пенроуза существенным образом является теоремой обычной математики (точнее – обычной математической логики, см. выше), но не конструктивной или интуиционистской математики. Если некто критикует теорему Пенроуза на основании того, что она использует метод доказательства от противного, то он примыкает к лагерю конструктивистов, и, будучи последовательным, должен отбросить существенную часть всей современной математики (в частности, почти весь математический анализ).

Из теоремы Гёделя–Тьюринга следует также, что мозг человека, способного понять теорему Гёделя–Тьюринга (назовем такого человека математиком), сам не является конечным автоматом. Действительно, предположим, что мозг математика – это некоторый конечный автомат. Тогда, используя теорему Гёделя–Тьюринга, математик может построить истинное утверждение, истинность которого он не может установить с использованием его собственного мозга как конечного автомата. Тем самым такой математик понимает то, что он понять не может. Противоречие доказывает, что исходное предположение было неверным и мозг математика не является конечным автоматом. Следовательно, аналогия «мозг – это компьютер» неверна для мозга любого отдельно взятого математика. Это утверждение можно рассматривать как альтернативную формулировку теоремы Пенроуза.

Интересно, что этим способом невозможно доказать, что мозг любого наперед заданного человека не является конечным автоматом. Теорема справедлива только либо в отношении математической способности всех людей, вместе взятых (среди которых есть, разумеется, и математики), либо в отношении отдельных личностей, понимающих или способных понять теорему Гёделя–Тьюринга. Но дальше логика Роджера Пенроуза такова. Математические способности людей представляют только частный случай способностей мозга, отличающийся от других способностей тем, что здесь анализ соотношения способностей мозга и компьютера удается абсолютно строго довести до конца. Поскольку в отношении математических способностей точно доказано превосходство человека над машиной, то, по аналогии, человек может обладать и другими способностями, недоступными конечному автомату. Просто анализ в других случаях довести до конца труднее, и, скорее всего, источник всех этих «невычислимых» способностей один. Действительно, многое указывает на существование таких способностей, и Роджер Пенроуз приводит множество примеров, причем не он первый это делает. Качественный анализ этого типа проводился и задолго до книг Пенроуза, например в книге Х. Дрейфуса [25] (конец 60-х гг. XX в.).

Таким образом, человек – не компьютер не только в смысле его математических способностей (что фактически доказано), но и во многих других отношениях. Проще говоря, самое главное, что недоступно компьютеру, это чисто человеческая универсальная способность к пониманию и придумыванию, которая, действительно, упорно не поддается формализации. Как написали братья Стругацкие в повести «Беспокойство»: «…все фундаментальные идеи выдумываются... они не висят на концах логических цепочек». Теорема Пенроуза указывает на то, что эта универсальная способность людей выдумывать не может быть формализована на основе архитектуры конечного автомата. Автоматы ходят только вдоль логических цепочек, поэтому новые фундаментальные идеи, которые не висят на их концах, им недостижимы. Но где они висят, эти фундаментальные идеи, как люди до них добираются? Мы этого не знаем.

Чтобы проиллюстрировать силу теоремы Пенроуза, приведем пару примеров, имеющих вид возражений против этой теоремы, и ответов на них.

Возражение 1. Если внимательно посмотреть на способ доказательства теоремы Гёделя–Тьюринга, то можно усмотреть, что способ генерации гёделевского утверждения сам имеет «механический» характер и может быть алгоритмизован. Пусть алгоритм, который генерирует гёделевское утверждение, включен в описание суперкомпьютера, о котором идет речь в доказательстве теоремы Пенроуза. Тогда он сможет самостоятельно сгенерировать гёделевское утверждение, якобы доступное только людям, но не этому компьютеру.

Ответ. Даже если алгоритм генерации гёделевских утверждений будет включен в описание суперкомпьютера, суперкомпьютер останется конечным автоматом, и на него по-прежнему будет распространяться теорема Гёделя–Тьюринга. Следовательно, и для обновленной конфигурации можно построить новое гёделевское утверждение, которое ему будет недоступно. Новое утверждение снова может быть включено в описание компьютера и т.д. Даже при неограниченной цепочке таких итераций компьютер останется конечным автоматом, следовательно, снова найдется хотя бы одно гёделевское утверждение, которое ему недоступно, но которое понимает математик-человек (по причине конструктивности этого утверждения).

Возражение 2. Выпишем все возможные способы математических рассуждений людей и заставим компьютер прочитать этот список. Тогда, очевидно, компьютеру станет доступно все, что доступно и людям.

Ответ. Выше было показано, что человек-математик не является конечным автоматом, следовательно, не является конечным автоматом и вся совокупность людей. Поэтому нельзя ожидать, что существует какой-то конечный список способов математических рассуждений, доступный людям. Список, вообще говоря, бесконечен, поэтому даже не может быть подготовлен для последующего прочтения.

В «Тенях разума» Роджер Пенроуз разбирает еще 18 разных возражений. Из двух возражений, разобранных выше, приведенный выше ответ на первое возражение следует Роджеру Пенроузу, на второе возражение мы здесь ответили по-другому.

Надо отметить, что анализ книги Рэя Курцвейла о технологической сингулярности [1] не идет дальше теоремы Гёделя–Тьюринга. Имя Пенроуза в контексте теоремы Гёделя–Тьюринга здесь даже не упоминается, и это при том, что с книгами Роджера Пенроуза Рэй Курцвейл точно знаком (что видно по обильному цитированию по другим поводам в других местах книги [1]). Это означает, что теорема Пенроуза просто не была понята. Например, в разделе «The Criticism from the Church-Turing Thesis» в книге [1] можно встретить следующее рассуждение (не дословно): компьютер – конечный автомат, поэтому для него существуют гёделевские утверждения. Это – несомненно. Но и мозг – тоже конечный автомат (это для Рэя Курцвейла ясно), поэтому и для него тоже существуют гёделевские утверждения. В этом

смысле компьютеры ничуть не хуже мозга, поэтому нет ничего удивительного в том, что компьютер может полностью воспроизвести работу мозга.

Следующий шаг от теоремы Гёделя–Тьюринга, который приводит к выводу о том, что мозг не может быть конечным автоматом, Рэй Курцвейл не сделал сам и не увидел его у Роджера Пенроуза.

Напротив, в книге Хьюберта Дрейфуса «Чего не могут вычислительные машины» [25] представлено ясное понимание того, что мозг не может быть просто компьютером. Собственно, в этом и состоит основной смысл книги. Интересен ответ на вопрос о том, почему мозг может не быть компьютером. Хьюберт Дрейфус формулирует его примерно так: машины обрабатывают информацию, а человек работает со смыслами. Вовсе не очевидно, что человеческие смыслы, чем бы они ни были на самом деле, могут быть закодированы информацией. Поэтому, возможно, мысль – не вычисление в обычном смысле. Как мы увидим ниже, Дрейфус мог оказаться очень близок к истине.

5. Природа невычислительной активности мозга по Роджеру Пенроузу

Тот тип способностей человека, который выходит за пределы вычислительных возможностей любого конечного автомата и существование которого доказывается теоремой Пенроуза, сам он называет невычислительной или невычислимой активностью мозга. Следующий вопрос, который обсуждает Пенроуз, состоит в следующем: что же в мозге человека содержится такого, что превращает его во что-то существенно отличное от любого вычислительного устройства в обычном понимании? Если мозг – не компьютер, то что это? В чем физически заключается корень невычислительной активности мозга?

Прежде всего Пенроуз показывает, что архитектура мозга как нейронной сети не имеет отношения к делу. Этот аспект проблемы связан с различием «нисходящих» и «восходящих» способов программирования. Под нисходящими способами понимаются традиционные методы, основанные на процедурном программировании, когда сначала пишется головная процедура программы, она обращается к подпрограммам, те обращаются к подпрограммам более низкого уровня и т.д., пока алгоритм не будет полностью детализирован вплоть до уровня машинных команд, не требующих дальнейшей детализации. Большинство программных систем до сих пор создаются именно этим способом. Восходящее программирование характерно для систем, в которых реализуется та или иная форма «обучения», и, на первый взгляд, какой-либо заранее определенный жесткий алгоритм, предназначенный для решения определенной группы задач, отсутствует. Наиболее типичным примером этого типа программирования являются искусственные нейронные сети. Однако отсутствие жесткого алгоритма в системе в этом случае является иллюзией. В действительности любая такая система может быть разложена на данные, которые могут представлять, например, структуру нейронной сети вместе с ее текущим состоянием и жесткое ядро алго-

ритма, который обрабатывает эти данные. Ядро управляет «обучением» системы и ее последующей «работой», просто модифицируя эти данные и интерпретируя их специальным способом. На фундаментальном теоретико-алгоритмическом уровне это есть просто особый сорт обработки данных конечным автоматом, и само ядро такой самообучающейся системы всегда имеет самую обыкновенную жесткую структуру, запрограммированную нисходящим способом. Такое ядро иногда может быть реализовано не программным путем, а «в железе», но это совершенно не меняет сути дела. В любой самообучающейся искусственной системе (из числа известных в настоящее время) может быть найдено жесткое алгоритмическое ядро. Так что использование искусственных нейронных сетей какой угодно сложности и любых других методов восходящего программирования вовсе не выводит за пределы понятия конечного автомата, и нейросетевая структура мозга, взятая сама по себе, тоже не имеет отношения к его невычислительной активности.

Значит, причину невычислительной активности мозга надо искать в чем-то другом. Логически остаются две возможности: либо этим чем-то может быть особая физика, управляющая работой мозга, либо источник невычислительной активности лежит вне мозга и соответствующая способность мозга является следствием того, что мозг является открытой системой. Обе эти возможности рассматриваются Роджером Пенроузом (то, что способность к невычислительной активности имеет «сверхъестественное» или «нематериальное» происхождение, Пенроуз не рассматривает в качестве варианта, который имеет смысл обсуждать).

На идею о том, что причина невычислительной активности мозга может быть следствием открытости мозга как системы и может лежать за пределами самого мозга, Пенроуз возражает двумя способами. Во-первых, если только предполагать, что любой конечный фрагмент окружения человека описывается вычислимой физикой, то он, формально говоря, может быть отображен на работу конечного автомата, поэтому ничего невычислимого в нем не найдется. В этом смысле невычислимости вне мозга просто неоткуда взяться. Никакие контакты мозга с окружением к невычислимости тогда не приведут. Во-вторых, если мы предполагаем, что вне человека всё-таки есть что-то, имеющее невычислимый характер, то мы, очевидно, соглашаемся, что невычислимые процессы в природе существуют, хотя бы в принципе. Но тогда нелепо предполагать, что они локализуются в каком-то месте, отличном от самого мозга, так как именно мозг является самой сложной из известных нам природных систем. То есть источник невычислительной активности мозга нужно искать только внутри мозга, никак не снаружи.

Далее Пенроуз детально рассматривает разные типы физических процессов в качестве потенциальной основы невычислительной активности внутри мозга (этому посвящена вторая половина каждой из книг [2; 3]). Его цель – найти невычислимую физику, так как такая физика автоматически могла бы привести и к невычислимому поведению мозга. Пенроуз приводит

несколько примеров, из которых следует, что в такой физике в принципе нет ничего невозможного. Например, если бы некоторая теория существенным образом использовала перечисление всех возможных топологий некоторого многомерного многообразия (например, в качестве индекса при суммировании какого-то ряда), то такая теория была бы невычислимой, так как доказано, что задача перечисления всех таких топологий алгоритмически неразрешима. Пенроуз рассматривает и другие «игрушечные» примеры невычислимой динамики.

Затем Пенроуз тщательно пересматривает в поисках невычислимости всю известную физику. Сначала он исключает классическую механику и классическую теорию поля, включая теорию гравитации (общую теорию относительности, ОТО). Все эти теории имеют простую динамическую природу и математически представляются либо задачей Коши для систем дифференциальных уравнений (обыкновенных или в частных производных), либо вариационными задачами для хорошо определенных функционалов. Здесь нет места невычислимости. Затем он рассматривает классическую статистическую физику. Для систем статистической физики не требуется вычислять частные динамические траектории, но требуется вычислять типичное поведение. Иными словами, задача сводится к моделированию ансамблей. Это снова не приводит к каким-либо принципиальным проблемам, связанным с вычислимостью. Ансамбли можно моделировать с использованием программных генераторов случайных чисел, которые могут следовать чисто случайному поведению с практически сколь угодно высокой точностью. Таким образом, вся классическая физика, как обыкновенная, так и статистическая, исключается полностью.

На этом этапе возникает, ни много ни мало, непрямое доказательство того, что мозг в процессе мышления существенным образом использует квантовые процессы. Действительно, если вся классическая физика, как источник невычислительной активности мозга, исключена, то остается искать такую активность в квантовой физике (как минимум). Это очень сильный вывод, который находится в глубоком противоречии с прогнозами Рэя Курцвейла и других приверженцев появления сильного ИИ в обозримом будущем. Такое доказательство существования квантовой активности мозга несколько напоминает доказательство существования гравитационного излучения по ускорению взаимного вращения в системах двойных пульсаров. Наблюдаемое ускорение находится в точном соответствии с предсказанием ОТО, основанном на квадрупольном излучении гравитационных волн, но сами гравитационные волны до сих пор не обнаружены в прямых экспериментах, поэтому нет полной уверенности в правильности интерпретации явления. В случае квантовой активности мозга тоже требуется ее прямое обнаружение, одного только доказательства через теорему Пенроуза не достаточно.

После классической физики Роджер Пенроуз рассматривает квантовую физику и снова не обнаруживает здесь ничего невычислимого. Этот его вы-

вод мы будем подробно обсуждать ниже, здесь же пока примем его и проследим дальнейшую логику. Получается, что никакая вообще известная физика к невычислимому поведению, которое демонстрирует мозг, привести не может, так как классическая и квантовая физика вместе исчерпывают вообще всю известную физику. Даже если мозг – не классический, а квантовый компьютер в современном понимании этого термина, это никак не может объяснить его невычислительную активность. Мозгу мало быть просто квантовым компьютером, нужно нечто большее. В основе работы мозга должна лежать еще неизвестная невычислимая физика, ничего другого не остается! Таков вывод Роджера Пенроуза. Сам Пенроуз спекулирует о том, что эта неизвестная физика может быть напрямую связана с квантовой гравитацией, которая до сих пор является камнем преткновения фундаментальной науки. То есть мозг является не квантовым компьютером, но «квантово-гравитационным компьютером».

Этот вывод Роджера Пенроуза понимается очень плохо, как и теорема Пенроуза об ИИ. В частности, неверное понимание продемонстрировано в книге Рэя Курцвейла [1]. Курцвейл предполагает, что Пенроуз считает мозг квантовым компьютером в обычном современном понимании, и с этой позиции Курцвейл разворачивает свою критику (кстати, тоже не вполне адекватную, о чем мы уже упоминали в параграфе 3). Истинная позиция Пенроуза осталась Курцвейлом непонятой, так как осталась непонятой и вся линия его рассуждений, начиная с доказательства no-go теоремы об ИИ.

Окончательный вывод Роджера Пенроуза о необходимости новой физики для понимания работы мозга кажется более чем поразительным. Получается, что в новую физику (квантовую гравитацию?) может вести не только создание гигантских коллайдеров для исследования физики микрочастиц или лишь ненамного менее гигантских и сложных телескопов для исследований в области космологии ранней Вселенной, гамма-барстеров и др., но наука о сознании. Мозг может быть таким же пробным камнем в поисках фундаментальной физики, как упомянутые ускорители и телескопы. И это есть мнение вполне прагматически и рационально настроенного ученого с мировым именем, сэра Роджера Пенроуза, но отнюдь не мистика-эзотерика. Надо, конечно же, отметить, что Пенроуз среди ученых вовсе не одинок в предположении, что наука о сознании, так или иначе, должна занять важное место в фундаментальной объединяющей физической теории. Хорошо известны исследования М.Б. Менского в этом направлении [26; 27; 28], недвусмысленно в этом же духе высказался Андрей Линде в недавнем интервью по случаю получения им премии Fundamental Physics Prize Юрия Мильнера [29]. Хотя, быть может, вывод Пенроуза и не столь уж удивителен. Действительно, физика элементарных частиц адресует экстремально мелкие пространственные масштабы, космология – Вселенную в целом, то есть экстремально большие масштабы, а мозг представляет самую сложную известную структуру – то есть экстремальную сложность.

Однако если вернуться к аргументам и выводам Роджера Пенроуза, касающимся интерпретации его no-go теоремы об ИИ, здесь хотелось бы дать некоторые уточнения, дополнения и поставить несколько вопросов. Они рассмотрены в следующих разделах статьи.

6. Невычислительная активность Вселенной в целом

Первое уточнение имеет довольно тривиальный характер и относится к утверждению Роджера Пенроуза, что вне мозга не надо искать невычислительную активность. Невычислительная активность, существующая вне мозга и не связанная прямо с чьей-либо еще мыслительной активностью, очевидным образом существует. Это активность Вселенной в целом. Активность Вселенной в целом принципиально не может быть «вычислена» конечным автоматом, так как для этого потребовался бы автомат (компьютер), заведомо превосходящий Вселенную в размере, что невозможно, так как он сам должен быть частью Вселенной. Активность Вселенной в целом невычислима.

Однако может ли этот сорт невычислимости быть причиной невычислительной активности мозга типа той, о которой идет речь в теореме Пенроуза? Вообще говоря, для сознания может быть существенно, что Вселенная, в силу своей практически актуальной бесконечности и невычислимости, является неисчерпаемой и для познания. Мы постоянно ожидаем и должны ожидать от природы чего-то, что абсолютно невозможно предсказать. Это обстоятельство вполне может каким-то образом стимулировать мыслительную активность и влиять на формы деятельности мозга. Но вряд ли бесконечность и невычислимость этого сорта имеет отношение к способности математика построить невыводимое для конечного автомата утверждение с использованием гёделевской нумерации и диагонального метода Кантора. В этой способности не просматривается какое-либо влияние актуальной бесконечности Вселенной. Хотя строго доказать, что такая связь отсутствует, вряд ли возможно, но она выглядит очень неправдоподобной или, как минимум, не первостепенной. Поэтому, как нам представляется, хоть аргумент о невычислимости бесконечной Вселенной и надо иметь в виду, но он не может повлиять на выводы Роджера Пенроуза.

7. Квантовая реальность: вычислимость и симулируемость

7.1. Вычислимость квантовой теории – «наивный» подход

Вывод Роджера Пенроуза о том, что вся классическая физика, включая классическую теорию поля и все разделы классической статистической физики, является существенно вычислимой и не может быть источником невычислительной активности мозга, представляется нам совершенно несомненным, и мы не будем здесь его подробно обсуждать. За деталями отсылаем к

книгам Пенроуза [2; 3]. На первый взгляд, вывод о том, что вся квантовая физика является вычислимой, тоже не вызывает никаких сомнений.

Действительно, состояния квантовых систем представляются векторами в гильбертовом пространстве. Вектор в гильбертовом пространстве – хорошо понятный объект, который в принципе может быть представлен в компьютере с любой заданной точностью, даже в том случае, если гильбертово пространство бесконечномерно. В принципе здесь ситуация весьма напоминает представление в компьютере обычных непрерывных классических полей с необходимой заданной точностью. С формальной точки зрения, вектор гильбертова пространства для компьютера – это просто цепочка комплексных чисел.

Эволюция систем в квантовой теории представляется унитарными преобразованиями состояний систем – векторов гильбертова пространства. С формальной точки зрения, такое унитарное преобразование есть либо умножение вектор-столбца, представляющего состояние системы, на унитарную матрицу, являющуюся преобразованием, отвечающим эволюции, либо результат решения некоторого линейного дифференциального уравнения с начальными условиями. В проведении таких операций для обычного компьютера нет ничего невозможного.

Наконец, третьим и последним компонентом квантовой теории являются измерения. Измерения характеризуются вероятностями получения на выходе измерительной процедуры тех или иных значений наблюдаемых величин, а сами вероятности определяются проекционным постулатом Борна – фон Неймана. Для вычисления вероятностей не требуется вычислять ничего более сложного, чем скалярные произведения векторов гильбертова пространства, и это, опять-таки, является алгоритмической процедурой, легко выполнимой на компьютере. Эти вероятности можно не только вычислить, но, при необходимости, даже симулировать соответствующее вероятностное поведение исхода измерения с использованием датчиков случайных чисел.

Таким образом, все вычисления, необходимые для предсказания поведения квантовых систем, формально говоря, можно выполнить на обычном классическом компьютере. Например, по этой причине можно поведение квантового компьютера полностью симулировать классическим конечным автоматом. Более того, такие программы-симуляторы существуют [30]. Философски говоря, квантовая реальность допускает исчерпывающее представление в классических вычислительных системах – компьютерах. Более точно можно говорить о принципиальной возможности изоморфизма фрагментов квантовой реальности и компьютерных (вычислительных) моделей этих фрагментов реальности.

7.2. Изоморфизм квантовой реальности и ее симуляции классическим конечным автоматом

Сформулированное выше представление о вычислимости квантовой теории приводит к некоторым не совсем тривиальным вопросам. Если кван-

товый процесс можно взаимно однозначно отобразить на функционирование классического конечного автомата, то, с точностью до изоморфизма, квантовый процесс – это просто и есть определенный способ работы конечного автомата. Но куда же тогда исчезает вся таинственность квантового мира? Где квантовые парадоксы, о которых столько сказано? Конечный автомат сводит квантовое поведение к чему-то простому и классическому, в чем ничего таинственного нет. Программа для компьютера не может быть таинственной. Более того, не имеем ли мы противоречия с доказанными теоремами о невозможности существования классических скрытых параметров в квантовой механике?

Действительно, любой классический компьютер представляет собой устройство, описываемое локальными параметрами, которые можно рассматривать и как скрытые параметры по отношению к моделируемой системе. Ячейки памяти – триггеры компьютера – представляют собой обычные классические системы, обладающие четкой локализацией в пространстве и времени. Состояния триггеров тоже полностью локализованы в пространстве-времени и соответствуют представлению о локальных классических переменных. И тем не менее, система, составленная из таких классических устройств, сама являющаяся классической и локальной, демонстрирует квантовое поведение. Как такое может быть? Что здесь не так?

Для того чтобы разобраться с этим вопросом, полезно иметь перед глазами пример простейшей нетривиальной системы, поведение которой, как это принято считать, прямо противоречит наличию классических скрытых параметров. Это ЭПР-пара скоррелированных частиц со спином 1/2. Для такой системы справедлива теорема Белла [31], которая говорит о том, что если в основе поведения системы лежат локальные скрытые параметры, то при измерении проекций спинов частиц должны выполняться определенные неравенства, которые называются неравенствами Белла. Реальные же квантовые ЭПР-пары нарушают эти неравенства, что прямо противоречит существованию локальных классических скрытых параметров.

Для дальнейшего обсуждения необходимы некоторые детали. В упомянутом эксперименте в начальном состоянии готовится пара частиц A и B со спинами 1/2 так, что спины частиц направлены точно в противоположные стороны, поэтому полный спин пары частиц равен нулю. Ничто другое про состояние системы неизвестно; такое состояние называется спиновым синглетом. После приготовления частицы могут быть разведены на произвольно большое расстояние или не разведены вовсе – это не имеет принципиального значения. Сам эксперимент состоит в том, что измеряются проекции спинов частиц A и B на различные направления. Для каждой частицы каждое измерение имеет всего два возможных исхода – либо спин направлен по выбранному направлению, либо против (это связано с тем, что спин частиц равен 1/2. Для больших значений спина количество исходов было бы больше). Если спин направлен по выбранному направлению, то результату измерения приписывается +1, если против, то –1. Измеряются корреляции результатов

измерений направлений спинов частиц *A* и *B* на разные направления. В данном случае эти корреляции представляют собой не что иное, как средние значения произведений результатов измерений.

Пусть для измерения проекций спинов используются две системы координат (*X,Y,Z*) и (*X',Y',Z'*), причем система координат (*X',Y',Z'*) повернута относительно системы (*X,Y,Z*) на угол Θ против часовой стрелки вокруг оси *Y*, а направления осей *Y* и *Y'* совпадают. Пусть S_Z^A означает результат измерения спина частицы *A* вдоль направления *Z* и т. д. Тогда теорема Белла утверждает, что если результаты измерений определяются локальными скрытыми параметрами, то должно выполняться неравенство:

$$|C| = |\langle S_Z^A S_{Z'}^B \rangle + \langle S_X^A S_{X'}^B \rangle + \langle S_X^A S_{Z'}^B \rangle - \langle S_Z^A S_{X'}^B \rangle| \leq 2, \quad (4)$$

где угловые скобки означают усреднение (корреляцию результатов измерений). Квантовая механика предписывает для величины *C* в уравнении (4) следующий результат:

$$C(\Theta) = 2\sqrt{2}\ \sin(\Theta - \pi/4). \quad (5)$$

Очевидно, что при некоторых углах Θ величина *C*(Θ) по модулю превышает двойку – неравенство (4) нарушается.

То, что классический компьютер вычислимым образом предсказывает нарушение неравенства Белла для ЭПР-пары, ясно уже из формулы (5). Эта формула является простым результатом применения проекционного постулата, и, конечно же, для компьютера не составляет проблемы вычислить значение корреляции для любого угла Θ с использованием этой формулы. Однако в связи с проблемой ИИ представляет интерес вопрос, может ли искусственное устройство симулировать работу мозга. Поэтому нас будет интересовать не просто вычислимость результата квантовой теории, но можно ли компьютер заставить работать как квантовую систему, то есть не предсказывать, а симулировать поведение квантовой системы. Ведь и теорема Белла говорит о том, что именно поведение системы с локальными скрытыми переменными не может нарушать неравенство [4].

Чтобы понять в деталях, как нарушение неравенства Белла может быть согласовано с поведением обычного классического компьютера, мы написали простую программу, которая должна симулировать поведение скоррелированной ЭПР-пары. Логически программа представляет собой структуру клиент-сервер (рис. 2). Сервером является та часть программы, в которой хранится состояние ЭПР-пары и в которой физически реализованы процедуры, связанные с преобразованием состояния при проведении некоторого измерения над этим состоянием. Два клиента *A* и *B* представляют измерения спинов, проводимые над частицами *A* и *B*. Технические подробности реализации сервера, «клиентов», и связи между ними мы опускаем, так как эти детали не играют роли в обсуждении.

Рис. 2. Структура программы, симулирующей измерение спинов над скоррелированной ЭПР-парой

Программа работает следующим образом. Сначала приготовляется начальное синглетное состояние ЭПР-пары, которое записывается в сервер, и сервер посылает сигнал готовности состояния в «клиенты». Это действие симулирует достижение частицами измерительных устройств. Получив сигнал, каждое из устройств должно провести измерение над состоянием объединенной системы двух частиц, которое хранится в сервере в единственном экземпляре. Очевидно, клиент должен обратиться к серверу за состоянием. Для этого каждый из клиентов посылает в сервер сигнал-запрос на измерение состояния частицы. Этот сигнал содержит информацию о том, для какой частицы и вдоль какой оси проводится измерение спина. Сервер обрабатывает этот сигнал в соответствии с проекционным постулатом. Для текущего состояния системы, записанного в сервере, он, используя датчик случайных чисел, генерирует результат измерения и, в соответствии с полученным результатом, проводит необходимое преобразование состояния системы (проектирование). Затем сервер возвращает результат измерения в тот клиент, который выдал запрос на измерение. Это есть завершение процедуры измерения спина – измерительное устройство получает результат измерения. Для каждого приготовленного состояния ЭПР-пары каждый клиент-измеритель обращается к серверу независимо от другого. Процедура повторяется многократно, так накапливается статистика для вычисления корреляций результатов измерений.

Сервер может обрабатывать запросы клиентов *A* и *B* только по очереди, так как сервер содержит единственный экземпляр состояния системы. Если программа правильно симулирует поведение ЭПР-пары, то результат симуляции (корреляции) не должен зависеть от порядка обработки запросов клиентов, так как в реальности этой зависимости нет и, более того, этот порядок, вообще говоря, даже не имеет физического смысла, так как измерения спинов могут быть разделены пространственно-подобным интервалом и иметь разную последовательность во времени для разных систем отсчета. Для того чтобы проверить независимость результата от порядка измерений, в программе предусмотрены три режима работы: сначала *A*, потом *B*; сначала *B*, потом *A*; случайный порядок. И, разумеется, никакой зависимости результатов измерений от их порядка, действительно, нет (вплоть до воспроизведения точного распределения статистики результатов измерений корреляций, если использовать одну и ту же последовательность случайных чисел).

В отношении измерения корреляций компьютерная модель ведет себя в точности как реальная ЭПР-пара (рис. 3). Неравенство Белла нарушено. Результаты «измерений» следуют предсказанной квантовой кривой [5] с ожидаемыми статистическими флуктуациями, связанными с ограниченным числом симуляций. Всё как в натурном эксперименте. Возникает два вопроса.

Q1 Получили ли мы противоречие с теоремой Белла, симулировав поведение квантовой скоррелированной пары классическим локальным устройством?

Q2 Получили ли мы на самом деле исчерпывающую симуляцию ЭПР-пары?

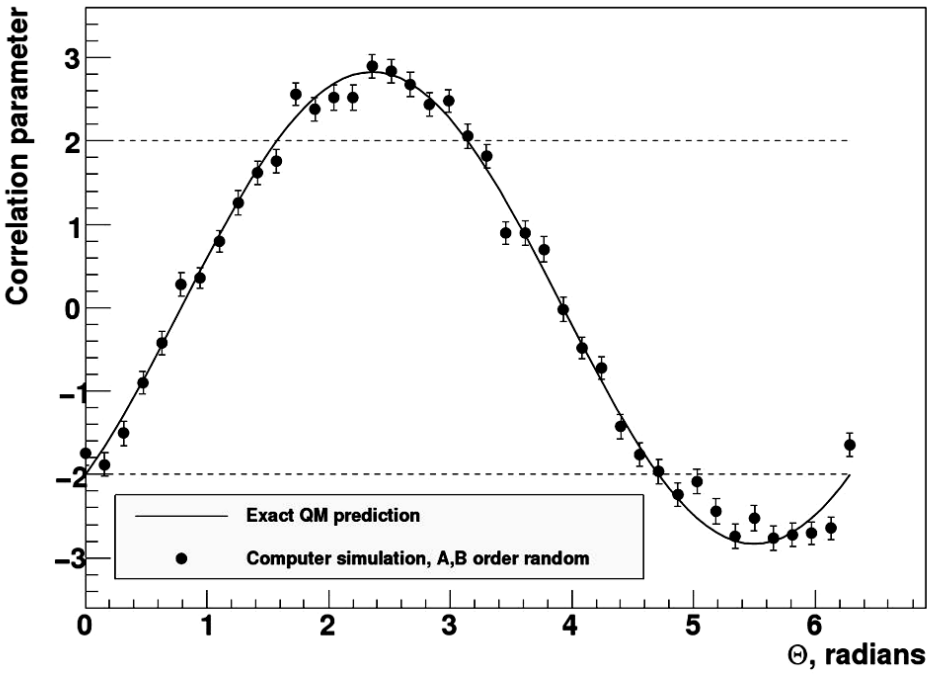

Рис. 3. Результат симуляции поведения ЭПР-пары компьютерной программой. Для каждого значения угла было симулировано по 100 измерений значений спинов. Горизонтальные пунктирные прямые представляют ограничения на величину корреляций, следующие из неравенства Белла. Видно, что неравенство Белла нарушено для некоторых углов и находится в точном соответствии с предсказанием

На вопрос Q1 можно уверенно дать отрицательный ответ. Никаких противоречий здесь нет. Всё дело в том, как в точности понимает локальные классические скрытые параметры теорема Белла. В доказательстве теоремы Белла (см. [31]) центральным пунктом является предположение, что акты измерения могут быть разделены пространственно-подобным интервалом, следовательно, они не могут оказывать причинное влияние друг на друга. Пусть a и b – направления, вдоль которых проводится измерение спина частиц, соответственно A и B. Тогда из причинной независимости результатов измерений следует, что направление b не может оказать влияние на резуль-

тат измерения A, и наоборот. Технически это соответствует тому, что при наличии локальных скрытых параметров в вычислении корреляций, как бы они ни считались, измерение A будет представлено некоторой функцией $A(a,\lambda)$, а измерение B – некоторой функцией $B(b,\lambda)$, но никак не функциями $A(a,b,\lambda)$ и $B(a,b,\lambda)$. Здесь λ обозначает скрытые переменные. В компьютерной модели роль скрытых переменных λ играет генератор случайных чисел. В нашей компьютерной модели это условие белловской локальности определенно не выполняется. Когда сервер выполняет запрос клиента A на измерение, он модифицирует состояние системы способом, зависимым от направления a, вдоль которого проводилось измерение спина A. Соответственно, когда обрабатывается последующий запрос на измерение B, то в состоянии системы уже зафиксировано влияние направления a, поэтому измерению B будет соответствовать функция вида $B(a,b,\lambda)$, но не $B(b,\lambda)$, как того требует теорема Белла. Акты измерения в компьютерной модели ЭПР-пары оказываются причинно-зависимыми, что не соответствует представлению о локальных скрытых переменных теоремы Белла. Таким образом, теорема Белла не запрещает существование классических скрытых параметров как таковых, но запрещает существование скрытых параметров, локализованных в областях, разделенных пространственно-подобными интервалами. Поэтому существование локальной классической компьютерной модели, нарушающей неравенства Белла, не противоречит теореме Белла.

Заметим, что понятие вычисления или эквивалентное ему понятие машины Тьюринга вообще никаким способом не адресует какие-либо пространственно-временные соотношения (кроме наивного представления о причинной связи последовательных шагов любого вычисления), поэтому заранее можно было бы догадаться, что теорема Белла не может иметь никакого отношения к функционированию какого-либо вычислительного устройства в обычном понимании. Однако пройденный нами путь полезен, так как приводит к детальному пониманию того, как именно и в каком смысле классическое устройство может симулировать «квантовую реальность». И, кстати, здесь немедленно возникает новый вопрос. Раз пространственно-временные связи и соотношения вообще нерелевантны понятию конечного автомата, но, очевидным образом, имеют самое прямое отношение к поведению реальных квантовых объектов вроде ЭПР-пар, разнесенных на пространственно-подобные интервалы, то, значит, представленная вычислительная реализация «квантовой реальности» упускает что-то очень существенное. Что именно?

Действительно, в реальном пространстве-времени вычислительными средствами симулировать пространственно-временное поведение квантовых систем, вообще говоря, невозможно (при том, что нарушение неравенства Белла возможно). Это легко понять, если сравнить пространственно-временную диаграмму поведения нашей компьютерной программы, симулирующей ЭПР-пару, и пространственно-временную диаграмму для реальной ЭПР-пары (рис. 4).

Представим себе, что система клиент-сервер, представленная на рис. 2, реально разнесена в пространстве, и расстояние между сервером (размерами которого мы пренебрегаем) и клиентами, представляющими измерения, одинаково для обоих клиентов и равно L. Никаких проблем в практической реализации такой системы нет. Для сравнения рассматриваем измерение спинов над ЭПР-парой на том же расстоянии L от места формирования исходного синглетного состояния пары. На обеих диаграммах начало координат соответствует пространственно-временной точке формирования ЭПР-пары (или состояния этой пары в компьютере). Для простоты будем предполагать, что все сигналы и частицы – компоненты ЭПР-пары – перемещаются с максимально возможной скоростью – со скоростью света. Начальная часть обеих диаграмм (стрелки из начала координат налево и направо вверх) представляет перемещение частиц от места рождения до приборов (реальная ЭПР-пара) и передачу сигнала готовности к измерителям (симуляция). В случае реальной ЭПР-пары измерение спинов происходит «мгновенно» (в пределе длительность может быть как угодно мала по сравнению с временами перемещения частиц). Напротив, в компьютерной симуляции время измерения принципиально конечно: «клиент-измеритель» должен отправить запрос к «серверу-состоянию», сервер должен отправить ответ «клиенту» и «клиент» должен ответ получить. Ответ на вопрос Q2 отрицательный: исчерпывающую симуляцию ЭПР-пары мы не получили. Пространственно-временные соотношения, характерные для реальной ЭПР-пары, в нашей модели нарушены.

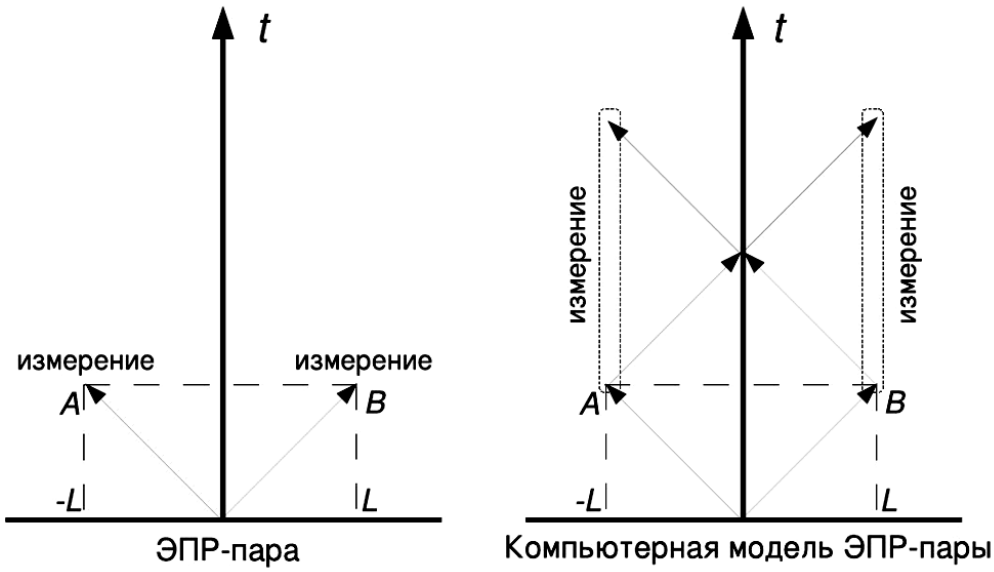

Рис. 4. Пространственно-временные диаграммы измерений спинов для реальной ЭПР-пары и для компьютерной системы, симулирующей измерения

Хотя была рассмотрена очень частная компьютерная модель, результат является вполне общим: никакая (распределенная) компьютерная система в реальном пространстве-времени в общем случае не может симулировать по-

ведение распределенной квантовой системы. То есть симулируемость поведения квантовых систем в реальном пространстве-времени не сводится к вычислимости квантовой теории.

Но что же тогда означает «полная вычислимость квантовой теории», о которой говорилось в 7.1, в отношении возможности симуляции квантовых систем, распределенных в пространстве-времени? Ответ очень простой: для полной симуляции таких систем само пространство-время принципиально может быть только виртуальным, симулированным. Невозможно классическую вычислительную систему заставить вести себя в реальном пространстве полностью как существенно распределенный в пространстве квантовый объект.

Покажем, как должна была бы быть устроена компьютерная программа, исчерпывающим образом симулирующая поведение ЭПР-пары, включая и все пространственно-временные связи. В такой модели, прежде всего, должна быть реализована модель реального (3+1)-мерного пространства-времени, в котором существуют приборы, измеряющие спины частиц, при этом сами классические измерительные приборы должны быть описаны достаточно детально (магнитное поле в установке Штерна–Герлаха, экраны или детекторы, фиксирующие появление частиц). Кроме того, модель должна содержать описание 6-мерного конфигурационного пространства (с учетом, кроме того, и спиновых переменных, на чем мы детально не останавливаемся), в котором совершают движение скоррелированные частицы ЭПР-пары (фактически – 6-мерный волновой пакет объединенного перепутанного состояния двух частиц). Заметим, что это 6-мерное пространство – не то же самое, что симулированное физическое пространство[19]. В этом контексте должно решаться уравнение Шредингера, описывающее распространение 6-мерного волнового пакета в конфигурационном пространстве. Когда по этому решению понятно, что частицы достигли зоны экранов или детекторов, применяется проекционный постулат, который порождает ответ в виртуальном трехмерном пространстве, моделирующем реальное пространство. Если использовалась установка типа Штерна–Герлаха, то проекционный постулат даст координату частицы на экране, по которой определяется и значение

[19] Полезно заметить, что даже при рассмотрении движения единственной точечной квантовой частицы, когда конфигурационное пространство является трехмерным, это конфигурационное пространство – не то же самое, что физическое трехмерное пространство. Это лишь частный случай общего 3N мерного конфигурационного пространства для N частиц при N=1. Волновая функция даже единственной точечной частицы не является волной, распространяющейся в физическом пространстве, как это часто наивно представляется! Поэтому нет никакого парадокса в том, что при измерении координат и сопровождающем его коллапсе, волновая функция частицы в конфигурационном пространстве схлопывается с бесконечной скоростью: это движение не происходит в физическом пространстве, где сверхсветовые скорости запрещены. Схлопывание происходит в конфигурационном пространстве, где скорость в обычном смысле вообще нельзя определить, так как мы не можем проникнуть в него с жесткими линейками и часами. «Появление прибора» в таком пространстве приводит к тому, что пространство становится другим (содержит описание степеней свободы прибора).

проекции спина (если частица ушла в одну сторону, то проекция спина +1, если в противоположную сторону, то –1). Измеренные значения спинов привязаны к точкам, разделенным пространственно-подобным интервалом виртуального пространства. Хотя логически состояние объединенной системы частиц и в такой модели должно храниться на некотором единственном локальном центральном сервере, но здесь не требуется никаких временных задержек (точнее, симуляции задержек) на общение между клиентами (приборами) и сервером, так как в виртуальном пространстве можно смоделировать бесконечно быстрое обращение к серверу. Мы не связаны предельной скоростью распространения сигнала, которая имеет отношение только к реальному пространству. Симуляция ЭПР-пары получается полной, но не в реальном, а в виртуальном пространстве-времени.

Мозг, если он действительно опирается в своей работе на квантовые процессы (что кажется неизбежным по теореме Пенроуза), вполне может являться распределенной квантовой системой, подобной ЭПР-паре (но в миллиарды раз сложнее). Поэтому для того, чтобы говорить о вычислимости квантовых процессов мозга, надо представлять себе реализацию модели мозга не в реальном, а в виртуальном пространстве-времени, подобно тому как это было описано выше для ЭПР-пары. Но при этом модель мозга должна обрабатывать сигналы из реального пространства времени и «жить» в реальном пространстве-времени, чтобы иметь какую-то связь с действительностью. Как это возможно? Для такого мозга должно быть два пространства-времени – внутреннее, виртуальное, в котором моделируются квантовые корреляции, и внешнее, реальное, в котором мозг «живет». Не возникнет ли неустранимых противоречий в попытке совместить эти две вещи в одной модели? Например, какое пространство-время должны адресовать модели внутренних сенсоров мозга, сигнализирующих о головной боли (и должна ли у компьютерной модели мозга болеть голова, если она представляет мозг, страдающий головной болью?). Как классические подсистемы мозга (вроде внешнего поведения нейронной сети как системы классических пороговых переключателей) в модели должны взаимодействовать с виртуальным пространством, в котором описываются корреляции? Впрочем, как мы видели, все эти вопросы не уникальны именно для модели квантового поведения мозга. Вопрос о виртуальном пространстве возникал уже при попытке симуляции нервной системы C. elegans на чисто классическом уровне (см. 2.2). Для квантового мозга все эти вопросы только многократно усложняются.

7.3. Неинформационная природа квантовых состояний

Как уже упоминалось, Хьюберт Дрейфус писал [25], что смыслы, которыми оперирует человеческое сознание, могут не иметь простой информационной природы. Как такое может случиться, Дрейфус не объяснил, поэтому это предположение осталось в рамках его книги абстрактной возможностью. Однако нетрудно понять, как такое может быть на самом деле.

Из теоремы Пенроуза и анализа вычислимости классической физики следует, что в классической (неквантовой) физике источник невычислимой активности мозга найти невозможно. Следовательно, приходится предполагать, что в активности мозга существенную роль играют, как минимум, обычные квантовые процессы (если не квантово-гравитационные, как предполагает Пенроуз). С этой логикой довольно трудно спорить. Далее, так как невычислительная активность мозга существенным образом сказывается на поведении человека, что опять-таки следует из теоремы Пенроуза, то следует ожидать, что квантовые процессы играют существенную роль в формировании смыслов, которые и определяют поведение.

Центральным действующим лицом любого квантового процесса является квантовое состояние – именно оно изменяется со временем. Но квантовое состояние не кодирует информацию. Действительно, одним из важнейших свойств информации является то, что ее можно «прочитать» и создать с использованием этой прочитанной информации любое число копий оригинала. С квантовым состоянием это сделать принципиально невозможно. Существует так называемая теорема о невозможности квантового клонирования (см. напр. [27. С. 301–304]), которая говорит о том, что заранее неизвестное квантовое состояние не может быть скопировано на другую систему без разрушения состояния исходной системы. Квантовые состояния нельзя копировать, тем самым квантовые состояния не обладают одним из важнейших свойств, которым обязана обладать информация, и сами не кодируют информацию.

При этом есть нечто, присущее квантовым состояниям, что очень напоминает информацию. Одним из важнейших свойств обычной информации является то, что одна и та же информация может быть представлена разными способами, на разных носителях. Точно так же квантовое состояние, имеющее одну и ту же структуру, могут иметь квантовые системы совершенно разной природы. Например, состояние двухуровневой квантовой системы с амплитудами $\left(\sqrt{1/3},\ \sqrt{2/3}\right)$ может быть и спиновым состоянием относительно некоторого заданного направления, и состоянием частицы в двухъямном потенциале. Независимо от природы систем, их состояния здесь имеют что-то общее, подобно тому, как, например, текст на экране компьютера и лежащем рядом клочке бумажки может иметь одинаковое информационное содержание. То общее, что может объединять квантовые состояния систем разной природы, логично назвать квантовой информацией. Но термин не должен вводить в заблуждение: квантовая информация не является информацией в обычном смысле, так как принципиально не допускает копирования.

Коль скоро мозг существенным образом использует квантовые процессы для формирования и обработки смыслов, то и смыслы, о которых писал Дрейфус, по крайней мере частично, связаны с квантовыми состояниями и не имеют простой информационной природы. Скорее всего, смыслы могут иметь смешанную информационно-квантовоинформационную природу.

Обычные компьютеры обрабатывают только информацию в обычном понимании, поэтому мозг, обрабатывая и квантовую информацию, уже только по этой причине не является обычным компьютером.

Квантовые состояния нельзя копировать, но можно переносить в единственном экземпляре с одной системы на другую похожую систему (с той же размерностью гильбертова пространства состояний). Такой перенос можно осуществить с помощью процесса квантовой телепортации [27. С. 309–315]. В частности, если в основе феномена сознания лежат квантовые процессы, то состояние сознания принципиально невозможно «скопировать» на искусственный носитель, оставив «оригинал» сознания в целости и сохранности. Это запрещает теорема о невозможности квантового клонирования. Можно сделать лишь одно из двух. Либо скопировать «устройство» мозга на искусственный носитель, но при этом будет скопирована лишь мертвая оболочка сознания, не само сознание. Либо можно телепортировать сознание на правильно подготовленный искусственный носитель, но тогда оригинальный носитель сознания автоматически окажется в «бессознательном» состоянии, так как все квантовые состояния, важные для содержания сознания, в нем будут разрушены. Надо также заметить, что процесс такой телепортации представляется невероятно сложным, единственная ошибка испортит все дело (проблема в квантовой когерентности), поэтому вряд ли такой процесс удастся осуществить в обозримом будущем.

7.4. Парадокс квантовой информации и «чудо клонирования»

Представление о неинформационной природе квантового состояния немедленно приводит к новой загадке. Действительно, в реальном мире квантовое состояние нельзя скопировать никакими силами. Однако, предположим, мы создали точную вычислительную модель некоторой квантовой системы, включая симуляцию физического пространства, в котором «живет» квантовая система. Это возможно в силу вычислимости квантовой теории. Как уже говорилось, квантовая система и ее вычислительная модель в классическом компьютере – это, с точностью до изоморфизма, одно и то же. Но в вычислительной модели, в отличие от реальной квантовой системы, квантовые состояния уже имеют ясную информационную природу. Это просто последовательности чисел, кодированные битами компьютерной памяти, которые представляют векторы гильбертова пространства (см. 7.1). Очевидно, что можно создать сколько угодно копий такого объекта (хотя копии невозможно будет создать только средствами, формализованными внутри модели). Почему в компьютерной симуляции квантовое состояние представляет собой информацию, а в оригинале квантовой системы – нет? Не противоречит ли обычная информационная природа симулированного квантового состояния неинформационной природе реальных квантовых состояний?

Нам представляется, что противоречие, действительно, есть и оно очень серьезно. По нашему мнению, наличие этого противоречия означает, что оригинальная квантовая система и ее даже очень совершенная вычислитель-

ная модель на классическом компьютере, – в действительности не одно и то же с точностью до изоморфизма, в отличие от того, что предполагалось при постановке вопроса (предыдущий абзац). Мы ясно понимаем, что возможен также иной взгляд на вещи. А именно можно заключить, что информационная природа компьютерной модели квантового состояния указывает на то, что и реальные квантовые состояния имеют информационную природу, несмотря на существование теоремы о запрете клонирования. Нам, однако, представляется, что эта точка зрения неверна. Она неверна, в частности, потому, что существенно обедняет понятие квантовой реальности. Однако, и мы должны это явно признать, мы не можем более точно сформулировать, в чем же состоит порок упомянутого выше «изоморфизма»; порок настолько серьезный, что отношение между даже самыми простыми квантовыми системами и их совершенными компьютерными моделями нельзя считать тождеством с точностью до изоморфизма.

Можно также заметить, что квантовое состояние и симуляция квантового состояния классическим компьютером представляют разные типы реальности. Как подробно обсуждается в нашей статье [24], достаточным критерием объективного существования некоторого объекта является возможность получения о нем воспроизводимых данных воспроизводимыми методами. Одиночное квантовое состояние не удовлетворяет этому критерию: воспроизводимая методика измерений (например измерение спина в установке Штерна–Герлаха) приводит, вообще говоря, к невоспроизводимым результатам (непредсказуемое направление проекции спина). Напротив, симуляция квантового состояния на классическом компьютере вполне удовлетворяет критерию объективной реальности. В отличие от одиночного квантового состояния, критерию объективной реальности удовлетворяет также ансамбль систем, представляющий квантовое состояние. Для ансамбля состояний воспроизводимые измерения приводят к воспроизводимым результатам в отношении различных распределений вероятности, а использование методов квантовой томографии позволяет полностью восстановить квантовое состояние, соответствующее ансамблю. Восстановив же это состояние, его можно скопировать на состояние другой квантовой системы с той же размерностью гильбертова пространства состояний или просто записать на классический носитель информации для последующего использования. То есть, в отличие от одиночного квантового состояния, ансамбль квантовых состояний является информационным понятием в том смысле, что в нем может быть закодирована информация в обычном смысле.

Парадоксальное противоречие между неинформационной природой квантового состояния и информационной природой симуляции квантового состояния можно превратить в еще более сильный парадокс. Предположим, что на классическом компьютере удалось очень точно смоделировать настолько большой фрагмент квантовой реальности, что он может включать в себя разумного наблюдателя. Если симуляция достаточно точна, то такой виртуальный наблюдатель не сможет заметить, что он живет в виртуальном,

а не в реальном мире. Но теперь мы в нашей компьютерной симуляции начнем создавать копии состояний некоторых квантовых систем, которые наблюдатель способен исследовать. Для нас («программистов») здесь нет никакой проблемы, так как в компьютерной модели квантовые состояния – это просто классические цепочки бит. Однако наблюдателю известна теорема о запрете клонирования состояний и он должен будет расценить «феномен клонирования», происходящий у него на глазах как чудо, нарушение законов природы. Более того, ничто не мешает в этой компьютерной симуляции реализовать модуль, который будет отслеживать состояние сознания наблюдателя (которое тоже есть не более, чем цепочка бит), и если в сознании будет обнаружено достаточно сильное желание обнаружить чудо клонирования, то этот программный модуль немедленно и представит его взору наблюдателя. Тогда, субъективно, наблюдатель будет сам вызывать чудо клонирования в ответ на «молитву».

Эта ситуация выглядит и вовсе абсурдной. В следующем параграфе мы покажем, как можно снять хотя бы сильную форму парадокса квантовой информации в форме «чуда клонирования».

7.5. Космологический горизонт вычислимости

Возможность разрешения парадокса «чуда клонирования» кроется в критическом анализе понятия вычислимости по отношению к квантовой теории. Фактически для вычисления поведения некоторых практически важных квантовых систем требуются такие мощности классического компьютера или такие объемы вычислений, которые нельзя реализовать не только практически, но и принципиально, так как для вычислений будет необходимо время, превышающее возраст Вселенной, или размер компьютера будет таков, что его нельзя будет уместить внутри космологического горизонта событий. Вот простейший пример. Для того чтобы с использованием квантового алгоритма Шора разложить на простые множители 1000-значное двоичное число (обычная задача для перспективных квантовых компьютеров), квантовому компьютеру требуется память всего в тысячу квантовых ячеек памяти – кубитов[20], в то время как классическому компьютеру для представления состояния такого квантового компьютера потребуется память $2^{1000} \approx 10^{300}$ комплексных чисел. Это на много порядков больше объема информации, которая может быть записана во всем обычном веществе видимой части Вселенной ($\sim 10^{90}$ бит, [32]). Даже если внутри видимого космологического горизонта попытаться разместить по одному биту информации в каждой планковской ячейке пространства (объемом порядка 10^{-99} см^3), то всего поместится порядка 10^{183} бит, что все еще почти на 120 порядков меньше необходимого. Квантовую систему, отвечающую 1000-кубитному квантовому компьютеру, принципиально невозможно симулировать на классическом компьютере, поэтому, в частности, реальные классические

[20] Фактически может потребоваться и несколько тысяч ячеек для реализации алгоритмов квантовой коррекции кода, но это несущественно для наших оценок.

компьютерные симуляторы квантовых вычислений могут работать только с очень малоразмерными системами.

Этот пример представляет очень общую проблему. Точное решение практически любой нетривиальной многочастичной проблемы в квантовой теории требует сверхкосмологически больших вычислительных ресурсов. По этой причине, например, принциально неразрешима проблема точного вычисления спектров возбуждений атомных ядер, состоящих больше чем из нескольких нуклонов и т.д.

Подчеркнем, что ограничение на возможность симуляции сложных (многочастичных) квантовых систем имеет принципиальный характер. Это ограничение связано, как мы видели, с фундаментальными характеристиками Вселенной, в которой мы обитаем (космологический горизонт событий, связанный с конечным возрастом Вселенной), и его уместно назвать космологическим горизонтом вычислимости. Он столь же фундаментален и непреодолим, как космологический горизонт событий. Понятие вычислимости в отношении сложных многочастичных квантовых систем лишено физического смысла, если необходимые ресурсы классического компьютера выходят за космологический горизонт вычислимости. Такие системы не могут быть исчерпывающим образом симулированы классическим компьютером в нашей Вселенной. Но, поскольку нам доступна единственная Вселенная, следует считать, что они вовсе не допускают симуляции классическими средствами, не являются вычислимыми в смысле космологического горизонта. Такая невычислимость имеет статус закона природы. Это очень существенная модификация «наивного» представления о вычислимости квантовой теории, представленного в 7.1.

Понятие космологического горизонта вычислимости позволяет снять парадокс «чуда клонирования». Любые достаточно большие фрагменты квантовой реальности, очевидным образом, невычислимы в смысле космологического горизонта. Симулированного наблюдателя невозможно поместить в квантовую реальность, симулированную на классическом компьютере так, чтобы он этого не заметил, даже в том случае, если сознание наблюдателя имеет чисто классическую природу. Внимательно изучив поведение многочастичных квантовых систем, симулированный наблюдатель обязательно обнаружит обман, так как точная симуляция таких систем классическим компьютером невозможна. А если сознание имеет квантовую природу, то тогда и просто адекватная симуляция самого наблюдателя будет невозможной. В любом случае, с использованием классического компьютера ситуацию, в которой имеет место «чудо клонирования», создать принципиально невозможно. В принципе большие фрагменты квантовой реальности могут быть симулированы квантовым компьютером, но в квантовом компьютере уже невозможно произвольное копирование информации в силу теоремы о запрете квантового клонирования и квантовой природы самого компьютера. Парадокс снимается полностью.

8. Космологический горизонт вычислимости, перспективы симуляции мышления и природа теоремы Пенроуза

Если работа мозга на некотором фундаментальном уровне опирается на что-то похожее на квантовые вычисления, то размерность соответствующей квантовой системы (или систем) будет столь высока, что ее физика заведомо будет невычислимой в смысле космологического горизонта вычислимости. Говорить о вычислимости поведения квантового мозга физически бессмысленно. Однако физика мозга все еще сохраняет свою алгоритмически разрешимую (вычислимую) природу в чисто математическом смысле. Вопрос состоит в том, должны ли мы считать квантовую активность мозга вычислимой или нет, если она невычислима с физической, но вычислима с математической точки зрения?

С нашей точки зрения, этот вопрос настолько сложен, что следует признать, что на него пока невозможно дать вполне определенный ответ. В то же время идея Роджера Пенроуза искать природу невычислимой активности мозга в еще неизвестной физике основана на определенном ответе «да» на этот вопрос. Представляется, что такая определенность является несколько преждевременной, поэтому возможность симулировать мышление человека не выходя за пределы квантовых вычислений не кажется закрытой.

Понятие космологического горизонта вычислимости позволяет несколько по-новому взглянуть на саму теорему Пенроуза. Доказательство Пенроуза апеллирует к теореме Гёделя–Тьюринга, а теорема Гёделя–Тьюринга использует представление о завершающихся и незавершающихся алгоритмах. Когда речь идет о завершающихся алгоритмах, то нигде в теореме, конечно, не затрагивается вопрос о том, как быстро они завершаются. Важна только «принципиальная» завершаемость. Роджер Пенроуз поясняет эту мысль следующим примером. Он рассматривает следующую задачу для компьютера: «распечатать последовательность $2^{2^{65563}}$ единиц, после чего остановиться» [3. С 141]. Пенроуз считает, что такую задачу следует считать завершающейся (вычислимой), несмотря на то, что необходимое время на такое вычисление превышает космологический горизонт вычислимости. Поэтому теорема Гёделя–Тьюринга, а следовательно, и теорема Пенроуза, явно оперируют вычислимостью и невычислимостью чисто математического типа, независимого от космологического горизонта вычислимости.

Здесь возникает одна тонкость. Объект, невычислимый в обычном смысле, является невычислимым и в смысле космологического горизонта, здесь нет проблем. Однако объект, вычислимый в чисто математическом смысле, может в смысле космологического горизонта быть невычислимым. Теорема Гёделя–Тьюринга, на которую опирается теорема Пенроуза, утверждает, что гёделевское утверждение для любого конечного автомата может быть построено в принципе (само оно вычислимо в математическом смысле), но, однако, не гарантирует, что такое построение окажется в рамках космологического горизонта вычислимости. Это является просто следствием

природы математики, в которую понятие космологического горизонта вычислимости не входит.

Между тем доказательство теоремы Пенроуза подразумевает, что для любого конечного автомата мы можем продемонстрировать гёделевское утверждение явно. Это означает, что мы имеем в руках утверждение, истинность которого нам ясна, но машина при попытке определить его истинность обязательно зациклится. Это и понимается под невычислимой природой нашего суждения. Однако реальность может быть такова, что гёделевское утверждение мы принципиально не можем «иметь в руках» по той причине, что возможности генерации этого утверждения окажутся за пределами космологического горизонта вычислимости. Может быть, это и не произойдет, но доказательство теоремы нам ничего в этом смысле не гарантирует. Таким образом, строго говоря, это суждение, вообще говоря, не может быть сформулировано в нашей Вселенной, а машину даже не удастся запустить для проверки истинности этого суждения. Поэтому мы не сможем продемонстрировать превосходство нашего мышления над машиной «практически», причем причины этой невозможности очень фундаментальны. Однако правдой остается то, что нам всегда известен путь, по которому нужно двигаться, чтобы такое суждение сгенерировать. Эквивалентно ли это тому, что нам указанное суждение доступно с точки зрения процедуры доказательства теорем Гёделя–Тьюринга и Пенроуза? Не означает ли это, что уже теорема Гёделя–Тьюринга лишена физического смысла? И допустимо ли понятие физического смысла распространять на математические теоремы? Но ведь вопрос о создании сильного ИИ – это практический, а не абстрактно-математический вопрос. Мы не готовы ответить на эти вопросы, они требуют дальнейших размышлений. Как нам представляется, самое сложное в этих вопросах – это понять, какой методикой надо вооружиться для ответа на них.

Резюме

Очевидным выводом из аргументации, приведенной в данной статье, является то, что основным препятствием на пути создания сильного ИИ является вовсе не недостаточная мощность современных компьютеров, а множество принципиальных нерешенных проблем, не все из которых даже удается аккуратно сформулировать. Причем проблемы эти относятся к самым разным научным направлениям, начиная с биологии клетки и кончая, как ни странно, космологией и фундаментальными вопросам, касающимися природы реальности. Последнее обстоятельство, впрочем, не является новостью, оно было прекрасно понято Роджером Пенроузом и отражено в его книгах. По мнению автора, число этих проблем столь велико, и они настолько трудны, что создание сильного ИИ в ближайшие десятилетия представляется очень маловероятным.

Что касается теоремы Пенроуза об искусственном интеллекте, то при всей ее невероятной силе, которую автор признает и всячески старался подчеркивать в статье, в ней остаются тонкие и не вполне ясные места. В частности, как нам представляется, следствия из теоремы не исключают однозначно возможность реализации сильного ИИ на основе только квантовых вычислительных устройств в обычном понимании, в отличие от того, что предполагает Роджер Пенроуз. Необходимость новой физики для понимания работы мозга не кажется пока совершенно неизбежной. Не вполне понятно также отношение природы теоремы Пенроуза к введенному в статье понятию космологического горизонта вычислимости. Все эти вопросы требуют дальнейшей работы. Интересно и полезно то, что анализ теоремы стимулирует новый взгляд на старые проблемы.

ЛИТЕРАТУРА

1. *Kurzweil R.* The singularity is near: when humans transcend biology / R. Kurzweil. – USA: Viking Penguin, 2005.
2. *Пенроуз Р.* Новый ум короля / Р. Пенроуз. – М.: УРСС, 2003.
3. *Пенроуз Р.* Тени разума: В поисках науки о сознании / Р. Пенроуз. – Москва-Ижевск: Институт компьютерных исследований, 2005.
4. Artificial intelligence (AI) // Encyclopedia Britannica. – Режим доступа: http://global.britannica.com/EBchecked/topic/37146/artificial-intelligence-AI
5. Searle J. Minds, Brains and Programs / J. Searle // Behavioral and Brain Sciences. – 1980. – Т. 3. – № 3. – С. 417–427.
6. *Vernor Vinge V.* The Coming Technological Singularity: How to Survive in the Post-Human Era / V. Vinge // VISION – 21 Symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute, March 30–31 – 1993. – Режим доступа: http://www.rohan.sdsu.edu/faculty/vinge/misc/singularity.html
7. *Joy B.* Why the future doesn't need us / B. Joy // Wired – April 2000. – P. 238–262.
8. *Forster H.* Doomsday: Friday, 13, November, A.D. 2026. / H. Forster, P. Mora, L. W. Amiot. // Science. – V. 132. – 1960. – P. 1291–1295.
9. *Шкловский И.С.* Вселенная, жизнь, разум / И.С. Шкловский. – М.: Наука, 1965.
10. Робот-учёный делает открытия без помощи человека. – Режим доступа: http://www.infuture.ru/article/1917
11. *Редько В.Г.* Эволюция, нейронные сети, интеллект. Модели и концепции эволюционной кибернетики / В.Г. Редько. – М.: Книжный дом «ЛИБРОКОМ», 2013.
12. *Дунин-Барковский. В.Л.* К вопросу об обратном конструировании мозга / В.Л. Дунин-Барковский // Глобальное будущее 2045. Конвергентные технологии и трансгуманистическая эволюция / под ред. проф. Д. И. Дубровского. – М.: ООО «Издательство МБА», 2013. – С. 150–157.
13. *Boyle J.H.* C. elegans locomotion: an integrated approach / J.H. Boyle. PhD thesis. – The University of Leeds School of Computing, 2009.
14. Слизевики – «разумные» простейшие. – Режим доступа: http://lostlab.ru/forum/topic432.html
15. *Александров В.Я.* Проблемы поведения на клеточном уровне – цитоэтология / В.Я. Александров // Успехи современной биологии. – 1970. – Т. 69. – № 2. – С. 220–240.
16. *Панов А.Д.* Универсальная эволюция и проблема поиска внеземного разума (SETI) / А.Д. Панов. – М.: ЛКИ (URSS), 2008.

17. *Режабек Б.Г.* О поведении механорецепторного нейрона в условиях замыкания его цепью искусственной обратной связи / Б. Г. Режабек // ДАН СССР. – 1971. – Т. 198 – № 4. – С. 981–984.
18. *Penrose R.* Consciousness in the Universe: Neuroscience, Quantum Space Time Geometry and Orch OR Theory / Roger Penrose, Stuart Hameroff // Journal of Cosmology. – Vol. 14. – 2011. Режим доступа: http://journalofcosmology.com/Consciousness160.html
19. Cytoskeleton // Encyclopedia Britannica. – Режим доступа: http://global.britannica.com/EBchecked/topic/148990/cytoskeleton
20. *Craddock T.J.A.* Cytoskeletal Signaling: Is Memory Encoded in Microtubule Lattices by CaMKII Phosphorylation? / T.J.A. Craddock, J.A. Tuszynski, S. Hameroff // PLOS Comput Biol. – V. 8 (3). – 2012. – e1002421.
21. *T. van der Sar T.* Decoherence-protected quantum gates for a hybrid solid-state spin register / T. van der Sar, Z. H. Wang, M. S. Blok, H. Bernien, T. H. Taminiau, D. M. Toyli, D. A. Lidar, D. D. Awschalom, R. Hanson, V. V. Dobrovitski // Nature. – 2012. – V. 484. – P. 82–86.
22. *Фейнман Р.* Фейнмановские лекции по физике. – Т. 8 / Р. Фейнман, Р. Лейнтон, М. Сэндс. – М.: Мир, 1966.
23. Клини С.К. Введение в метаматематику / С. К. Клини. – М.: Изд-во иностранной литературы, 1957.
24. *Панов А.Д.* Природа математики, космология и структура реальности: объективность мира математических форм / А.Д. Панов // Космология, физика, культура. – М.: ИФ-РАН, 2011 – С. 191–219.
25. *Дрейфус Х.* Чего не могут вычислительные машины / Х. Дрейфус. – Изд. 2-е. – М.: ЛИБРОКОМ (URSS), 2010.
26. *Менский М.Б.* Концепция сознания в контексте квантовой механики / М.Б. Менский // Успехи Физических Наук. – 2005. – Т. 175. – № 4. – С. 413–435.
27. *Менский М Б.* Человек и квантовый мир / М. Б. Менский. – Фрязино: Век 2, 2007.
28. *Менский М.Б.* Сознание и квантовая механика. Жизнь в параллельных мирах (Чудеса сознания – из квантовой реальности) / М.Б. Менский. – Фрязино: Век 2, 2011.
29. *Линде А.Д.* Интервью для журнала «Вокруг Света» (Сергей Добрынин) / А.Д. Линде // Вокруг Света. – 2012. – № 10 (2865). – С. 139–148.
30. *Julia-Daz B.* Qdensity – a Mathematica quantum computer simulation / B. Julia-Daz, J.M. Burdis, and F. Tabakin. // arXiv:quant-ph/0508101 – 2005.
31. *Гриб А.А.* Неравенства Белла и экспериментальная проверка квантовых корреляций на макроскопических расстояниях / А. А. Гриб // Успехи физических наук. – 1984. – Т. 142 – С. 619–634.
32. *Гуревич И.М.* Информация – всеобщее свойство материи: Характеристики, оценки, ограничения / И.М. Гуревич, У.А. Урсул. – М.: Книжный дом «ЛИБРОКОМ» (URSS), 2012.

THE TECHNOLOGICAL SINGULARITY, PENROSE THEOREM ABOUT ARTIFICIAL INTELLIGENCE AND QUANTUM NATURE OF CONSCIOSNESS

A.D. Panov

The predictions related to possibility of the strong artificial intelligence creation in the nearest decades are critically discussed. It is argued that the predictions are based on three weakly-based assumptions and on one completely misunderstood issue. The three weakly-based assumptions are: 1) the possibility of creation of strong artificial intelligence is determined by availability of sufficiently powerful computers; 2) the brain capacity is determined by the overall capacity of

synaptic connections; 3) the brain capacity can be estimated on the base of the analogy `brain is a classical computer'. The completely misunderstood issue is the arguments related to the Penrose theorem about artificial intelligence which forbids the creation of a computer in possession of all human capabilities on the base of the architecture of a finite machine. The critical analysis of all these three assumptions, the Penrose theorem, and the corollaries from the theorem deduced by Rodger Penrose himself is given. The Penrose' conclusions are estimated to be too pessimistic.

Key words: artificial intelligence, technological singularity, Goedel-Turing theorem, finite state machine, quantum computer, computability.