

Джон Сёрль

Разум мозга – компьютерная программа?

*Нет. Программа лишь манипулирует символами, мозг же придает им
смысл*

СПОСОБНА ли машина мыслить? Может ли машина иметь сознанные мысли в таком же смысле, в каком имеем их мы? Если под машиной понимать физическую систему, способную выполнять определенные функции (а что еще под ней можно понимать?), тогда люди – это машины особой, биологической разновидности, а люди могут мыслить, и, стало быть, машины, конечно, тоже могут мыслить. Тогда, по всей видимости, можно создавать мыслящие машины из самых разнообразных материалов – скажем, из кремниевых кристаллов или электронных ламп. Если это и окажется невозможным, то пока мы, конечно, этого еще не знаем.

Однако в последние десятилетия вопрос о том, может ли машина мыслить, приобрел совершенно другую интерпретацию. Он был подменен вопросом: способна ли машина мыслить только за счет выполнения заложенной в нее компьютерной программы? Является ли программа основой мышления? Это принципиально иной вопрос, потому что он не затрагивает физических, каузальных (причинных) свойств существующих или возможных физических систем, а скорее относится к абстрактным, вычислительным свойствам формализованных компьютерных программ, которые могут быть реализованы в любом материале, лишь бы он был способен выполнять эти программы.

Довольно большое число специалистов по искусственному интеллекту (ИИ) полагают, что на второй вопрос следует ответить положительно; другими словами, они считают, что составив правильные программы с правильными входами и выходами, они действительно создадут разум. Более того, они полагают, что имеют в своем распоряжении научный тест, с помощью которого можно судить об успехе или неудаче такой попытки. Имеется в виду тест Тьюринга, изобретенный Аланом М. Тьюрингом, основоположником искусственного интеллекта. Тест Тьюринга в том смысле, как его сейчас понимают, заключается просто в следующем: если компьютер способен демонстрировать поведение, которое эксперт не сможет отличить от поведения человека, обладающего определенными мыслительными способностями (скажем, способностью выполнять операции сложения или понимать китайский язык), то компьютер также обладает этими способностями. Следовательно, цель заключается лишь в том, чтобы создать программы, способные моделировать человеческое мышление таким образом, чтобы выдерживать тест Тьюринга. Более того, такая программа

будет не просто моделью разума; она в буквальном смысле слова сама и будет разумом, в том же смысле, в котором человеческий разум – это разум.

Конечно, далеко не каждый специалист по искусственному интеллекту разделяет такую крайнюю точку зрения. Более осторожный подход заключается в том, чтобы рассматривать компьютерные модели как полезное средство для изучения разума, подобно тому как они применяются при изучении погоды, пищеварения, экономики или механизмов молекулярной биологии. Чтобы провести различие между этими двумя подходами, я назову первый «сильным ИИ», а второй – «слабым ИИ». Важно понять, насколько радикальным является подход сильного ИИ. Сильный ИИ утверждает, что мышление – это не что иное, как манипулирование формализованными символами, а именно это и делает компьютер: он оперирует формализованными символами. Подобный взгляд часто суммируется примерно следующим высказыванием: «Разум по отношению к мозгу – это то же, что и программа по отношению к аппаратуре компьютера».

СИЛЬНЫЙ ИИ отличается от других теорий разума по крайней мере в двух отношениях: его можно четко сформулировать, но также четко и просто его можно опровергнуть. Характер этого опровержения таков, что каждый человек может попробовать провести его самостоятельно. Вот как это делается. Возьмем, например, какой-нибудь язык, которого вы не понимаете. Для меня таким языком является китайский. Текст, написанный по-китайски, я воспринимаю как набор бессмысленных каракулей. Теперь предположим, что меня поместили в комнату, в которой расставлены корзинки, полные китайских иероглифов. Предположим также, что мне дали учебник на английском языке, в котором приводятся правила сочетания символов китайского языка, причем правила эти можно применять, зная лишь форму символов, понимать значение символов совсем необязательно. Например, правила могут гласить: «Возьмите такой-то иероглиф из корзинки номер один и поместите его рядом с таким-то иероглифом из корзинки номер два».

Представим себе, что находящиеся за дверью комнаты люди, понимающие китайский язык, передают в комнату наборы символов и что в ответ я манипулирую символами согласно правилам и передаю обратно другие наборы символов. В данном случае книга правил есть не что иное, как «компьютерная программа». Люди, написавшие ее, – «программисты», а я играю роль «компьютера». Корзинки, наполненные символами, – это «база данных»; наборы символов, передаваемых в комнату, это «вопросы», а наборы, выходящие из комнаты, это «ответы».

Предположим далее, что книга правил написана так, что мои «ответы» на «вопросы» не отличаются от ответов человека, свободно владеющего китайским языком. Например, люди, находящиеся снаружи, могут передать непонятные мне символы, означающие; «Какой цвет вам больше всего нравится?» В ответ, выполнив предписанные правилами манипуляции, я выдам символы, к сожалению, мне также непонятные и означающие, что мой любимый цвет синий, но мне также очень нравится зеленый. Таким образом, я выдержу тест Тьюринга на понимание китайского языка. Но все же на

самом деле я не понимаю ни слова по-китайски. К тому же я никак не могу научиться этому языку в рассматриваемой системе, поскольку не существует никакого способа, с помощью которого я мог бы узнать смысл хотя бы одного символа. Подобно компьютеру, я манипулирую символами, но не могу придать им какого бы то ни было смысла.

Сущность этого мысленного эксперимента состоит в следующем: если я не могу понять китайского языка только потому, что выполняю компьютерную программу для понимания китайского, то и никакой другой цифровой компьютер не сможет его понять таким образом. Цифровые компьютеры просто манипулируют формальными символами согласно правилам, зафиксированным в программе.

То, что касается китайского языка, можно сказать и о других формах знания. Одного умения манипулировать символами еще недостаточно, чтобы гарантировать знание, восприятие, понимание, мышление и т. д. И поскольку компьютеры как таковые – это устройства, манипулирующие символами, наличия компьютерной программы недостаточно, чтобы можно было говорить о наличии знания.

Этот простой аргумент имеет решающее значение для опровержения концепции сильного ИИ. Первая предпосылка аргумента просто констатирует формальный характер компьютерной программы. Программы определяются в терминах манипулирования символами, а сами символы носят чисто формальный, или «синтаксический» характер. Между прочим, именно благодаря формальной природе программы, компьютер является таким мощным орудием. Одна и та же программа может выполняться на машинах самой различной природы, равно как одна и та же аппаратная система способна выполнять самые разнообразные компьютерные программы. Представим это соображение кратко в виде «аксиомы»:

Аксиома 1. Компьютерные программы – это формальные (синтаксические) объекты.

Это положение настолько важно, что его стоит рассмотреть несколько подробнее. Цифровой компьютер обрабатывает информацию, сначала кодируя ее в символических обозначениях, используемых в машине, а затем манипулируя символами в соответствии с набором строго определенных правил. Эти правила представляют собой программу. Например, в рамках тьюринговской концепции компьютера в роли символов выступали просто 0 и 1, а правила программы предписывали такие операции, как «Записать 0 на ленте, продвинуться на одну ячейку влево и стереть 1». Компьютеры обладают удивительным свойством: любая представимая на естественном языке информация может быть закодирована в такой системе обозначений и любая задача по обработке информации может быть решена путем применения правил, которые можно запрограммировать.

ВАЖНОЕ значение имеют еще два момента. Во-первых, символы и программы – это чисто абстрактные понятия: они не обладают физическими свойствами, с помощью которых их можно было бы определить и реализовать в какой бы то ни было физической среде. Нули и единицы, как

символы, не имеют физических свойств. Я акцентирую на этом внимание, поскольку иногда возникает соблазн отождествить компьютеры с той или иной конкретной технологией – скажем, с кремниевыми интегральными микросхемами – и считать, что речь идет о физических свойствах кремниевых кристаллов или что синтаксис означает какое-то физическое явление, обладающее, может быть, еще неизвестными каузальными свойствами аналогично реальным физическим явлениям, таким как электромагнитное излучение или атомы водорода, которые обладают физическими, каузальными свойствами. Второй момент заключается в том, что манипуляция символами осуществляется без всякой связи с каким бы то ни было смыслом. Символы в программе могут обозначать все, что угодно программисту или пользователю. В этом смысле программа обладает синтаксисом, но не обладает семантикой.

Следующая аксиома является простым напоминанием об очевидном факте, что мысли, восприятие, понимание и т. п. имеют смысловое содержание. Благодаря этому содержанию они могут служить отражением объектов и состояний реального мира. Если смысловое содержание связано с языком, то в дополнение к семантике, в нем будет присутствовать и синтаксис, однако лингвистическое понимание требует по крайней мере семантической основы. Если, например, я размышляю о последних президентских выборах, то мне в голову приходят определенные слова, но эти слова лишь потому относятся к выборам, что я придаю им специфическое смысловое значение в соответствии со своим знанием английского языка. В этом отношении они для меня принципиально отличаются от китайских иероглифов. Сформулируем это кратко в виде следующей аксиомы:

Аксиома 2. Человеческий разум оперирует смысловым содержанием (семантикой).

Теперь добавим еще один момент, который был продемонстрирован экспериментом с китайской комнатой. Располагать только символами как таковыми (т. е. синтаксисом) еще недостаточно для того, чтобы располагать семантикой. Простого манипулирования символами недостаточно, чтобы гарантировать знание их смыслового значения. Кратко представим это в виде аксиомы.

Аксиома 3. Синтаксис сам по себе не составляет семантику и его недостаточно для существования семантики.

На одном уровне этот принцип справедлив по определению. Конечно, кто-то может определить синтаксис и семантику по-иному. Главное, однако, в том, что существует различие между формальными элементами, не имеющими внутреннего смыслового значения, или содержания, и теми явлениями, у которых такое содержание есть. Из рассмотренных предпосылок следует:

Закключение 1. Программы не являются сущностью разума и их наличия недостаточно для наличия разума.

А это по существу означает, что утверждение сильного ИИ ложно.

Очень важно отдавать себе отчет в том, что именно было доказано с помощью этого рассуждения и что нет.

Во-первых, я не пытался доказывать, что «компьютер не может мыслить». Поскольку все, что поддается моделированию вычислениями, может быть описано как компьютер, и поскольку наш мозг на некоторых уровнях поддается моделированию, то отсюда тривиально следует, что наш мозг – это компьютер, и он, разумеется, способен мыслить. Однако из того факта, что систему можно моделировать посредством манипулирования символами и что она способна мыслить, вовсе не следует, что способность к мышлению эквивалентна способности к манипулированию формальными символами.

Во-вторых, я не пытался доказывать, что только системы биологической природы, подобные нашему мозгу, способны мыслить. В настоящее время это единственные известные нам системы, обладающие такой способностью, однако мы можем встретить во Вселенной и другие способные к осознанным мыслям системы, а может быть, мы даже сумеем искусственно создать мыслящие системы. Я считаю этот вопрос открытым для споров.

В-третьих, утверждение сильного ИИ заключается не в том, что компьютеры с правильными программами могут мыслить, что они могут обладать какими-то неизвестными доселе психологическими свойствами; скорее, оно состоит в том, что компьютеры просто должны мыслить, поскольку их работа – это и есть не что иное, как мышление.

В-четвертых, я попытался опровергнуть сильный ИИ, определенный именно таким образом. Я пытался доказать, что мышление не сводится к программам, потому что программа лишь манипулирует формальными символами – а, как нам известно, самого по себе манипулирования символами недостаточно, чтобы гарантировать наличие смысла. Это тот принцип, на котором основано рассуждение о китайской комнате.

Я подчеркиваю здесь эти моменты отчасти потому, что П.М. и П.С.Черчленды в своей статье (см. Пол М. Черчленд и Патриция Смит Черчленд «Может ли машина мыслить?»), как мне кажется, не совсем правильно поняли суть моих аргументов. По их мнению, сильный ИИ утверждает, что компьютеры в конечном итоге могут обрести способность к мышлению и что я отрицаю такую возможность, рассуждая лишь на уровне здравого смысла. Однако сильный ИИ утверждает другое, и мои доводы против не имеют ничего общего со здравым смыслом.

Далее я скажу еще кое-что об их возражениях. А пока я должен заметить, что в противоположность тому, что говорят Черчленды, рассуждение с китайской комнатой опровергает любые утверждения сильного ИИ относительно новых параллельных технологий, возникших под влиянием и моделирующих работу нейронных сетей. В отличие от компьютеров традиционной архитектуры фон Неймана, работающих в последовательном пошаговом режиме, эти системы располагают многочисленными вычислительными элементами, работающими параллельно и взаимодействующими друг с другом в соответствии с

правилами, основанными на открытиях нейробиологии. Хотя пока достигнуты скромные результаты, модели «параллельной распределенной обработки данных» или «коммутационные машины» подняли некоторые полезные вопросы относительно того, насколько сложными должны быть параллельные системы, подобные нашему мозгу, чтобы при их функционировании порождалось разумное поведение.

Однако параллельный, «подобный мозгу» характер обработки информации не является существенным для чисто вычислительных аспектов процесса. Любая функция, которая может быть вычислена на параллельной машине, будет вычислена и на последовательной. И действительно, ввиду того что параллельные машины еще редки, параллельные программы обычно все еще выполняются на традиционных последовательных машинах. Следовательно, параллельная обработка также не избегает аргумента, основанного на примере с китайской комнатой.

Более того, параллельные системы подвержены своей специфической версии первоначального опровергающего рассуждения в случае с китайской комнатой. Вместо китайской комнаты представьте себе китайский спортивный зал, заполненный большим числом людей, понимающих только английский язык. Эти люди будут выполнять те же самые операции, которые выполняются узлами и синапсами в машине коннекционной архитектуры, описанной Черчлендами, но результат будет тем же, что и в примере с одним человеком, который манипулирует символами согласно правилам, записанным в руководстве. Никто в зале не понимает ни слова по-китайски, и не существует способа, следуя которому вся система в целом могла бы узнать о смысловом значении хотя бы одного китайского слова. Тем не менее при правильных инструкциях эта система способна правильно отвечать на вопросы, сформулированные по-китайски.

У параллельных сетей, как я уже говорил, есть интересные свойства, благодаря которым они могут лучше моделировать мозговые процессы по сравнению с машинами с традиционной последовательной архитектурой. Однако преимущества параллельной архитектуры, существенные для слабого ИИ, не имеют никакого отношения к противопоставлению между аргументом, построенным на примере с китайской комнатой, и утверждением сильного ИИ. Черчленды упускают из виду этот момент, когда говорят, что достаточно большой китайский спортивный зал мог бы обладать более высокими умственными способностями, которые определяются размерами и степенью сложности системы, равно как и мозг в целом более «разумен», чем его отдельные нейроны. Возможно и так, но это не имеет никакого отношения к вычислительному процессу. С точки зрения выполнения вычислений последовательные и параллельные архитектуры совершенно идентичны: любое вычисление, которое может быть произведено в машине с параллельным режимом работы, может быть выполнено машиной с последовательной архитектурой. Если человек, находящийся в китайской комнате и производящий вычисления эквивалентен и той и другой системам, тогда, если он не понимает

китайского языка исключительно потому, что ничего кроме вычислений не делает, то и эти системы также не понимают китайского языка. Черчленды правы, когда говорят, что первоначальный довод, основанный на примере с китайской комнатой, был сформулирован исходя из традиционного представления об ИИ, но они заблуждаются, считая что параллельная архитектура делает этот довод неуязвимым. Это справедливо в отношении любой вычислительной системы. Производя только формальные операции с символами (т. е. вычисления) вы не сможете обогатить свой разум семантикой, независимо от того выполняются эти вычислительные операции последовательно или параллельно; вот почему аргумент китайской комнаты опровергает сильный ИИ в любой его форме.

МНОГИЕ люди, на которых этот аргумент производит определенное впечатление, тем не менее затрудняются провести четкое различие между людьми и компьютерами. Если люди, по крайней мере в тривиальном смысле, являются компьютерами и если люди обладают семантикой, то почему они не могут наделить семантикой и другие компьютеры? Почему мы не можем запрограммировать компьютеры Вах или Стау таким образом, чтобы у них тоже появились мысли и чувства? Или почему какая-нибудь новая компьютерная технология не сможет преодолеть пропасть, разделяющую форму и содержание, или синтаксис и семантику? В чем на самом деле состоит то различие между биологическим мозгом и компьютерной системой, благодаря которому аргумент с китайской комнатой действует применительно к компьютерам, но не действует применительно к мозгу?

Наиболее очевидное различие заключается в том, что процессы, которые определяют нечто как компьютер (а именно вычислительные процессы), на самом деле совершенно не зависят от какого бы то ни было конкретного типа аппаратной реализации. В принципе можно сделать компьютер из старых жестяных банок из-под пива, соединив их проволокой и обеспечив энергией от ветряных мельниц.

Однако когда мы имеем дело с мозгом, то хотя современная наука в значительной степени еще пребывает в неведении относительно протекающих в мозгу процессов, мы поражаемся чрезвычайной специфичности анатомии и физиологии. Там, где мы достигли некоторого понимания того, как мозговые процессы порождают те или иные психические явления, – например, боль, жажду, зрение, обоняние – нам ясно, что в этих процессах участвуют вполне определенные нейробиологические механизмы. Чувство жажды, по крайней мере в некоторых случаях, обусловлено срабатыванием нейронов определенных типов в гипоталамусе, которое в свою очередь вызвано действием специфического пептида, ангиотензина II. Причинные связи прослеживаются здесь «снизу вверх» в том смысле, что нейронные процессы низшего уровня обуславливают психические явления на более высоких уровнях. В самом деле, каждое «ментальное» явление, от чувства жажды до мыслей о математических

теоремах и воспоминаний о детстве, вызывается срабатыванием определенных нейронов в определенных нейронных структурах.

Однако почему эта специфичность имеет такое важное значение? В конце концов всевозможные срабатывания нейронов можно смоделировать на компьютерах, физические и химические свойства которых совершенно отличны от свойств мозга. Ответ состоит в том, что мозг не просто демонстрирует формальные процедуры или программы (он делает и это тоже), но и вызывает ментальные события благодаря специфическим нейробиологическим процессам. Мозг по сути своей является биологическим органом и именно его особые биохимические свойства позволяют достичь эффекта сознания и других видов ментальных явлений. Компьютерные модели мозговых процессов обеспечивают отражение лишь формальных аспектов этих процессов. Однако моделирование не следует смешивать с воспроизведением. Вычислительные модели ментальных процессов не ближе к реальности, чем вычислительные модели любого другого природного явления.

Можно представить себе компьютерную модель, отражающую воздействие пептидов на гипоталамус, которая будет точна вплоть до каждого отдельного синапса. Но с таким же успехом мы можем представить себе компьютерное моделирование процесса окисления углеводов в автомобильном двигателе или пищеварительного процесса в желудке. И модель процессов, протекающих в мозге, ничуть не реальнее моделей, описывающих процессы сгорания топлива или пищеварительные процессы. Если не говорить о чудесах, то вы не сможете привести свой автомобиль в движение, моделируя на компьютере окисление бензина, и вы не сможете переварить обед, выполняя программу, которая моделирует пищеварение. Представляется очевидным и тот факт, что и моделирование мышления также не произведет нейробиологического эффекта мышления.

Следовательно, все ментальные явления вызываются нейробиологическими процессами мозга. Представим сокращенно этот тезис следующим образом:

Аксиома 4. *Мозг порождает разум.*

В соответствии с рассуждениями, приведенными выше, я немедленно прихожу к тривиальному следствию.

Закключение 2. *Любая другая система, способная породить разум, должна обладать каузальными свойствами (по крайней мере), эквивалентными соответствующим свойствам мозга.*

Это равносильно, например, следующему утверждению: если электрический двигатель способен обеспечивать автомашине такую же высокую скорость, как двигатель внутреннего сгорания, то он должен обладать (по крайней мере) эквивалентной мощностью. В этом заключении ничего не говорится о механизмах. На самом деле, мышление — это биологическое явление: психические состояния и процессы обусловлены процессами мозга. Из этого еще не следует, что только биологическая система может мыслить, но это в то же время означает, что любая система

другой природы, основанная на кремниевых кристаллах, жестяных банках и т. п., должна будет обладать каузальными возможностями, эквивалентными соответствующим возможностям мозга. Таким образом, я прихожу к следующему выводу:

Заключение 3. Любой артефакт, порождающий ментальные явления, любой искусственный мозг должен иметь способность воспроизводить специфические каузальные свойства мозга, и наличия этих свойств невозможно добиться только за счет выполнения формальной программы.

Более того, я прихожу к важному выводу, касающемуся человеческого мозга:

Заключение 4. Тот способ, посредством которого человеческий мозг на самом деле порождает ментальные явления, не может сводиться лишь к выполнению компьютерной программы.

ВПЕРВЫЕ сравнение с китайской комнатой было приведено мною на страницах журнала "Behavioral and Brain Sciences" (Науки о поведении и мозге) в 1980 г. Тогда моя статья сопровождалась, в соответствии с принятой в этом журнале практикой, комментариями оппонентов, в данном случае свои соображения высказали 26 оппонентов. Откровенно говоря, мне кажется, что смысл этого сравнения довольно очевиден, но, к моему удивлению, статья и в дальнейшем вызвала целый поток возражений, и что еще более удивительно, этот поток продолжается и по сей день. Повидимому, аргумент китайской комнаты затронул какое-то очень болезненное место.

Основной тезис сильного ИИ заключается в том, что любая система (независимо, сделана ли она из пивных банок, кремниевых кристаллов или просто из бумаги) не только способна обладать мыслями и чувствами, но просто *должна* ими обладать, если только она реализует правильно составленную программу, с правильными входами и выходами. Очевидно, это абсолютно антибиологическая точка зрения, и естественно было бы ожидать, что специалисты по искусственному интеллекту охотно откажутся от нее. Многие из них, особенно представители молодого поколения, согласны со мной, но меня поражает, как много сторонников имеет эта точка зрения и как настойчиво они защищают ее. Приведу некоторые наиболее распространенные из высказываемых ими доводов:

а) В китайской комнате вы на самом деле понимаете китайский, хотя и не отдаете себе в этом отчета. В конце концов вы можете понимать что-то и не отдавая себе в этом отчета.

б) Вы не понимаете китайского, но в вас существует подсистема (подсознание), которая понимает. Существуют ведь подсознательные психические состояния, и нет причины считать, что ваше понимание китайского не могло бы быть полностью неосознанным.

в) Вы не понимаете китайского, но комната как целое – понимает. Вы подобны отдельному нейрону в мозгу, и нейрон как таковой не может ничего понимать, он лишь вносит свой вклад в то понимание, которое

демонстрирует система в целом; вы сами не понимаете, но вся система понимает.

г) Никакой семантики и не существует: есть только синтаксис. Полагать, что в мозгу есть какое-то загадочное «психическое содержание», «мыслительные процессы» или «семантика», это своего рода донаучная иллюзия. Все, что на самом деле существует в мозгу, – это некоторое синтаксическое манипулирование символами, которое осуществляется и в компьютерах. И ничего больше.

д) В действительности вы не выполняете компьютерную программу – это вам только кажется. Если существует некий сознательный агент, следующий строкам программы, то процесс уже вовсе не является простой реализацией программы.

е) Компьютеры обладали бы семантикой, а не только синтаксисом, если бы их входы и выходы были поставлены в причинные, каузальные зависимости – по отношению к остальному миру. Допустим, что мы снабдили робота компьютером, подключили телевизионные камеры к его голове, установили трансдюсеры, подводящие телевизионную информацию к компьютеру, и позволили последнему управлять руками и ногами робота. В таком случае система как целое будет обладать семантикой.

ж) Если программа моделирует поведение человека, говорящего по-китайски, то она понимает китайский язык. Предположим, что нам удалось смоделировать работу мозга китайца на уровне нейронов. Но тогда, конечно, подобная система будет понимать китайский так же хорошо, как и мозг любого китайца.


И так далее.

У всех этих доводов есть одно общее свойство: все они неадекватны рассматриваемой проблеме, потому что не улавливают самой сути рассуждения о китайской комнате. Эта суть заключается в различии между формальным манипулированием символами, осуществляемым компьютером, и смысловым содержанием, биологически порождаемым мозгом, – различии, которое я для краткости выражения (и надеюсь, что это никого не введет в заблуждение) свел к различию между синтаксисом и семантикой. Я не буду повторять своих ответов на все эти возражения, однако проясню, возможно, ситуацию, если скажу, в чем заключаются слабости наиболее распространенного довода моих оппонентов, а именно довода (в), который я назову ответом системы. (Очень часто встречается также и довод (ж), основанный на идее моделирования мозга, но он уже был рассмотрен выше.)

В ОТВЕТЕ системы утверждается, что *вы*, конечно, не понимаете китайского, но вся система в целом – вы сами, комната, свод правил, корзинки, наполненные символами, – понимает. Когда я впервые услышал это объяснение, я спросил высказавшего это объяснение человека: «Вы что же, считаете, что комната может понимать китайский язык?» Он ответил, да. Это, конечно, смелое утверждение, однако, помимо того что оно совершенно неправдоподобно, оно не состоятельно еще и с чисто логической точки зрения. Суть моего исходного аргумента была в том, что простое тасование

символов еще не обеспечивает доступа к пониманию смысла этих символов. Но это в той же мере касается комнаты в целом, как и находящегося в ней человека. В правоте моих слов можно убедиться, несколько расширив наш мысленный эксперимент. Представим себе, что я заучил наизусть содержимое корзинок и книги правил и что я провожу все вычисления в уме. Допустим даже, что я работаю не в комнате, а у всех на виду. В системе не осталось ничего такого, чего бы не было во мне самом, но поскольку я не понимаю китайского языка, не понимает его и система.

В своей статье мои оппоненты Черчленды используют одну из разновидностей ответа системы, придумав любопытную аналогию. Предположим, кто-то стал утверждать, что свет не может иметь электромагнитную природу, поскольку, когда человек перемещает магнит в темной комнате, мы не наблюдаем видимого светового излучения. Приведя этот пример, Черчленды спрашивают, а не является ли аргумент с китайской комнатой чем-то в том же роде? Не равносильно ли будет сказать, что когда вы манипулируете китайскими иероглифами в семантически темной комнате, в ней не возникает никакого просвета в понимании китайского языка? Но не может ли потом в ходе будущих исследований выясниться – так же, как было доказано, что свет все-таки целиком состоит из электромагнитного излучения, – что семантика целиком и полностью состоит из синтаксиса? Не является ли этот вопрос предметом дальнейшего научного изучения?

Аргументы, построенные на аналогиях, всегда очень уязвимы, поскольку, прежде чем аргумент станет состоятельным, необходимо еще убедиться, что две рассматриваемые ситуации действительно аналогичны. В данном случае, я думаю, что это не так. Объяснение света на основе электромагнитного излучения – это причинное рассуждение от начала и до конца. Это причинное объяснение физики  электромагнитных волн. Однако аналогия с формальными символами не состоятельна, поскольку формальные символы не имеют физических причинных свойств. Единственное, что во власти символов как таковых, – это вызвать следующий шаг в программе, которую выполняет работающая машина. И здесь не возникает никакой речи о дальнейших исследованиях, которым еще предстоит раскрыть доселе неизвестные физические причинные свойства нулей и единиц. Последние обладают лишь одним видом свойств – абстрактными вычислительными свойствами, которые уже хорошо изучены.

Черчленды говорят, что у них «напрашивается вопрос», когда я утверждаю, что интерпретированные формальные символы не идентичны смысловому содержанию. Да, я, конечно, не тратил много времени на доказательство, что это так, поскольку я считаю это логической истиной. Как и в случае с любой другой логической истиной, каждый может быстро убедиться, что она справедлива, поскольку, предположив обратное, сразу приходишь к противоречию. Попробуем провести такое доказательство. Предположим, что в китайской комнате имеет место какое-то скрытое понимание китайского языка. Что же может превратить процесс манипулирования синтаксическими элементами в специфично китайское

смысловое содержание? Подумав, я в конце концов пришел к выводу, что программисты должны были говорить по-китайски, коль скоро они сумели запрограммировать систему для обработки информации, представленной на китайском языке.

Хорошо. Но теперь представим себе, что надоело, сидя в китайской комнате, тасовать китайские (для меня бессмысленные) символы. Предположим, мне пришло в голову интерпретировать эти символы как обозначения ходов в шахматной игре. Какой семантикой теперь обладает система? Обладает ли она китайской семантикой или шахматной, или она обладает одновременно и той и другой? Предположим, что есть еще некая третья личность, наблюдающая за мной в окошко, и она решает, что мое манипулирование символами можно интерпретировать как предсказание курса акций на бирже. И так далее. Не существует предела количеству семантических интерпретаций, которое можно приписать символам, поскольку, я повторяю, символы носят чисто формальный характер. Они не содержат в себе внутренней семантики.

Можно ли каким-то образом спасти аналогию Черчлендов? Выше я сказал, что формальные символы не имеют каузальных свойств. Но, конечно, программа всегда выполняется той или иной конкретной аппаратурой, и эта аппаратура обладает своими специфическими физическими, каузальными свойствами. Любой реальный компьютер порождает различные физические явления. Мой компьютер, к примеру, выделяет тепло, производит монотонный шум и т. д. Существует ли здесь какое-либо строгое логическое доказательство, что компьютер не может производить аналогичным образом эффект сознания? Нет. В научном смысле об этом и речи быть не может, однако это совсем не то, что призвано опровергать рассуждение о китайской комнате, и не то, на чем будут настаивать сторонники сильного ИИ, поскольку любой производимый таким образом эффект будет достигать за счет физических свойств реализующей программу среды. Основное утверждение сильного ИИ заключается в том, что физические свойства реализующей среды не имеют никакого значения. Имеют значение лишь программы, а программы – это чисто формальные объекты.

Таким образом аналогия Черчлендов между синтаксисом и электромагнитным излучением наталкивается на дилемму: либо синтаксис следует понимать чисто формально, через его абстрактные математические свойства, либо нет. Если выбрать первую альтернативу, то аналогия становится несостоятельной, поскольку синтаксис, понимаемый таким образом, не имеет физических свойств. Если же, с другой стороны, рассматривать синтаксис в плоскости физических свойств реализующей среды, тогда аналогия действительно состоятельна, но она не имеет отношения к сильному ИИ.

ПОСКОЛЬКУ сделанные мною утверждения довольно очевидны – синтаксис это не то же самое, что семантика; мозговые процессы порождают психические явления – возникает вопрос, а как вообще возникла эта путаница? Кому могло прийти в голову, что компьютерное моделирование

ментального процесса полностью ему идентично? В конце концов весь смысл моделей заключается в том, что они улавливают лишь какую-то часть моделируемого явления и не затрагивают остального. Ведь никто не думает, что мы захотим поплавать в бассейне, наполненном шариками для пинг-понга, моделирующими молекулы воды. Можно ли тогда считать, что компьютерная модель мыслительных процессов на самом деле способна мыслить?

Отчасти эти недоразумения объясняются тем, что люди унаследовали некоторые положения бихевиористских психологических теорий прошлого поколения. Под тестом Тьюринга скрывается соблазн считать, что если нечто ведет себя так, как будто оно обладает ментальными процессами, то оно и на самом деле должно ими обладать. Частью ошибочной бихевиористской концепции было также и то, что психология, для того чтобы оставаться научной дисциплиной, должна ограничиваться изучением внешне наблюдаемого поведения. Парадоксально, но этот остаточный бихевиоризм связан с остаточным дуализмом. Никто не думает, что компьютерная модель пищеварения способна что-то переварить на самом деле, но там, где речь идет о мышлении, люди охотно верят в такие чудеса, потому что забывают о том, что разум – это такое же биологическое явление, как и пищеварение. По их мнению, разум – это нечто формальное и абстрактное, а вовсе не часть полужидкой субстанции, из которой состоит наш головной мозг. Полемическая литература по искусственному интеллекту обычно содержит нападки на то, что авторы называют дуализмом, но при этом они не замечают, как сами демонстрируют ярко выраженный дуализм, поскольку если не принять точку зрения, что разум совершенно не зависит от мозга или какой-либо другой физически специфической системы, то следует считать невозможным создание разума только за счет написания программ.

Исторически в странах Запада научные концепции, в которых люди рассматривались как часть обычного физического или биологического мира, часто встречали противодействие со стороны реакции. Идеям Коперника и Галилея противились, потому что они отрицали, что Земля является центром Вселенной. Против Дарвина выступали потому, что он утверждал, что люди произошли от низших животных. Сильный ИИ правильнее всего было бы рассматривать как одно из последних проявлений этой антинаучной традиции, так как он отрицает, что человеческий разум содержит что-то существенно физическое или биологическое. Согласно утверждениям сильного ИИ, разум не зависит от мозга. Он представляет собой компьютерную программу и по существу не связан ни с какой специфической аппаратурой.

Многие люди, сомневающиеся относительно физической значимости искусственного интеллекта, полагают, что компьютеры, может быть, и смогут понимать китайский язык или думать о числах, но принципиально не способны на проявления чисто человеческих свойств, а именно (и далее следует их излюбленная человеческая специфика): любовь, чувство юмора, тревога за судьбу постиндустриального общества в эпоху современного

капитализма и т. д. Но специалисты по ИИ справедливо настаивают, что эти возражения не корректны, что здесь как бы отодвигаются футбольные ворота. Если искусственное моделирование интеллекта окажется успешным, то психологические вопросы уже не имеют сколь-нибудь важного значения, В этом споре обе стороны не замечают различия между моделированием и воспроизведением. Пока речь идет о моделировании, то не стоит никакого труда запрограммировать мой компьютер, чтобы он напечатал: «Я люблю тебя, Сюзи»; «Ха-ха!» или «Я испытываю тревоги постиндустриального общества». Важно отдавать себе отчет в том, что моделирование – это не то же самое, что воспроизведение; и этот факт имеет такое же отношение к размышлениям об арифметике, как и к чувству тревоги. Дело не в том, что компьютер доходит только до центра поля и не доходит до ворот. Компьютер даже не трогается с места. Он просто не играет в эту игру.