

Джон СЕРЛ

СОЗНАНИЕ, МОЗГ И ПРОГРАММЫ[1]

Какую психологическую и философскую значимость следует нам приписать недавним усилиям по компьютерному моделированию познавательных способностей человека? Я считаю, что, отвечая на этот вопрос, полезно отличать “сильный” AI (как я это называю) от “слабого” или “осторожного” AI (Artificial Intelligence — Искусственный Интеллект). Согласно слабому AI, основная ценность компьютера в изучении сознания состоит в том, что он дает нам некий очень мощный инструмент. Например, он дает нам возможность более строгим и точным образом формулировать и проверять гипотезы. Согласно же сильному AI, компьютер — это не просто инструмент в исследовании сознания; компьютер, запрограммированный подходящим образом, на самом деле и есть некое сознание в том смысле, что можно буквально сказать, что при наличии подходящих программ компьютеры понимают, а также обладают другими когнитивными состояниями. Согласно сильному AI, поскольку снабженный программой компьютер обладает когнитивными состояниями, программы — не просто средства, позволяющие нам проверять психологические объяснения; сами программы суть объяснения.

У меня нет возражений против притязаний слабого AI, во всяком случае, в этой статье. Мое обсуждение здесь будет направлено на притязания, которые я определил здесь как притязания сильного AI, именно: на притязание, согласно которому подходящим образом запрограммированный компьютер буквально образом обладает когнитивными состояниями, и тем самым программы объясняют человеческое познание. Когда я далее буду упоминать AI, я буду иметь в виду сильный вариант, выраженный в этих двух притязаниях.

Я рассмотрю работу Роджера Шэнка и его коллег в Йейле (Schank and Abelson 1977), так как я знаком с ней больше, чем с другими подобными точками зрения, и потому что она представляет собой очень ясный пример того типа работ, которые я хотел бы рассмотреть. Но ничего в моем последующем изложении не зависит от деталей программ Шэнка. Те же аргументы приложимы и к SHRDLU Винограда (Winograd 1973), и к ELIZA Вейценбаума (Weizenbaum 1965), и, в сущности, к любому моделированию феноменов человеческой психики средствами машин Тьюринга.

Очень коротко, опуская разнообразные детали, можно описать программу Шэнка следующим образом: цель программы — смоделировать человеческую способность понимать рассказы. Для способности людей понимать рассказы характерно, что люди способны отвечать на вопросы о данном рассказе даже в тех случаях, когда даваемая ими информация не выражена в рассказе явным образом. Так, например, представьте, что вам дан следующий рассказ: “Человек зашел в ресторан и заказал гамбургер. Когда гамбургер подали, оказалось, что он подгорел, и человек в гневе покинул ресторан, не заплатив за гамбургер и не оставив чаевых”. И вот если вас спросят: “Съел ли человек гамбургер?”, вы, видимо, ответите: “Нет, не съел”. Точно так же, если вам предъявят следующий рассказ: “Человек зашел в ресторан и заказал гамбургер; когда гамбургер подали, он ему очень понравился; и покидая ресторан, он перед оплатой по счету дал официантке большие чаевые,” и спросят: “Съел ли человек свой гамбургер?”, вы, видимо, ответите: “Да, съел”. И вот машины Шэнка могут точно так же отвечать на вопросы о ресторанах. В этих целях они обладают неким “представлением” (“репрезентацией”) той информации о ресторанах,

какая бывает у людей и какая дает людям возможность отвечать на подобные вопросы, когда им предъявлен некий рассказ, вроде тех, что приведены выше. Когда машине предъявляют рассказ и затем задают вопрос, она распечатает такой ответ, какой мы ожидали бы от человека, которому предъявлен подобный рассказ. Приверженцы сильного AI утверждают, что в этой последовательности вопросов и ответов машина не только моделирует некую человеческую способность, но что, кроме того:

1. можно сказать буквально, что машина понимает рассказ и дает ответы на вопросы;
2. то, что делают машина и ее программа, объясняет человеческую способность понимать рассказ и отвечать на вопросы о нем.

Мне, однако, представляется, что работа Шэнка[2] никоим образом не подкрепляет ни одно из этих двух утверждений, и я сейчас попытаюсь показать это.

Любую теорию сознания можно проверить, например, так: задаться вопросом, что бы это означало, что мое сознание на самом деле функционирует в соответствии с теми принципами, о которых данная теория утверждает, что в соответствии с ними функционируют все сознания. Приложим этот тест к программе Шэнка с помощью следующего Gedankenexperiment[3]. Представим себе, что меня заперли в комнате и дали некий массивный текст на китайском языке. Представим себе далее, что я не знаю ни письменного, ни устного китайского языка (так оно и есть на самом деле) и что я не уверен даже, что распознал бы китайский письменный текст в качестве такового, сумев отличить его, скажем, от японского письменного текста или от каких-нибудь бессмысленных закорючек. Для меня китайское письмо как раз и представляет собой набор бессмысленных закорючек. Представим себе далее, что вслед за этой первой китайской рукописью мне дали вторую китайскую рукопись вместе с набором правил сопоставления второй рукописи с первой. Правила эти на английском языке, и я понимаю их, как понял бы любой другой носитель английского языка. Они дают мне возможность сопоставить один набор формальных символов со вторым набором формальных символов, и единственное, что значит здесь слово “формальный”, — то, что я могу распознавать символы только по их форме. Представьте себе теперь, что мне дали третью китайскую рукопись вместе с некоторыми инструкциями, — вновь на английском языке, — дающими мне возможность сопоставлять элементы этой третьей рукописи с первыми двумя, и эти правила учат меня, как в ответ на те или иные формальные символы из третьей рукописи выдавать определенные китайские символы, имеющие определенные очертания. Люди, которые дали мне все эти символы, называют первый текст “рукописью”, второй — “рассказом”, а третий — “вопросами”, но я всех этих названий не знаю. Кроме того, они называют символы, которые я выдаю в ответ на третий текст, “ответами на вопросы”, а набор правил на английском языке, который они дали мне, — “программой”. Чтобы слегка усложнить эту историю, вообразите себе, что эти люди также дали мне некие рассказы на английском языке, которые я понимаю, и они задают мне вопросы на английском языке об этих рассказах, и я выдаю им ответы на английском языке. Представьте себе также, что после некоторого промежутка времени я так набиваю руку в выполнении инструкций по манипулированию китайскими символами, а программисты так набивают руку в писании программ, что при взгляде со стороны — то есть с точки зрения какого-либо человека, находящегося вне комнаты, в которой я заперт, — мои ответы на вопросы абсолютно неотличимы от ответов настоящих носителей китайского языка. Никто не сможет сказать,

— если он видел только мои ответы, — что я ни слова не говорю по-китайски. Представим себе далее, что мои ответы на английские вопросы неотличимы от ответов, которые бы дали настоящие носители английского языка (как этого и следовало ожидать) — по той простой причине, что я и есть настоящий носитель английского языка. При взгляде со стороны, — с точки зрения какого-нибудь человека, читающего мои “ответы”, — ответы на китайские вопросы и ответы на английские вопросы равно хороши. Но в случае китайских ответов, в отличие от случая английских ответов, я выдаю ответы, манипулируя неинтерпретированными формальными символами. Что же касается китайского языка, я веду себя попросту как компьютер; совершаю вычислительные операции на формальном образе определенных элементах. В том, что касается китайского языка, я есть просто инстанция компьютерной программы.

И вот претензии сильного AI состоят в том, что программированный компьютер понимает рассказы и его программа в некотором смысле объясняет человеческое понимание. Но мы теперь можем рассмотреть эти претензии в свете нашего мысленного эксперимента.

1. Что касается первой претензии, мне кажется совершенно очевидным, что в данном примере я не понимаю ни одного слова в китайских рассказах. Мои входы и выходы неотличимы от входов и выходов носителя китайского языка, и я могу обладать какой угодно формальной программой, и все же я ничего не понимаю. По тем же самым основаниям компьютер Шэнка ничего не понимает ни в каких рассказах — в китайских, в английских, в каких угодно, поскольку в случае с китайскими рассказами компьютер — это я, а в случаях, где компьютер не есть я, он не обладает чем-то большим, чем я обладал в том случае, в котором я ничего не понимал.

2. Что касается второй претензии, — что программа объясняет человеческое понимание, — мы видим, что компьютер и его программа не дают достаточных условий понимания, поскольку компьютер и программа работают, а между тем понимания-то нет. Но, может быть, при этом создается хотя бы необходимое условие или делается существенный вклад в понимание? Одно из утверждений сторонников сильного AI состоит в том, что когда я понимаю некий рассказ на английском языке, я делаю в точности то же самое — или, быть может, почти то же самое, — что я делал, манипулируя китайскими символами. Случай с английским языком, когда я понимаю, отличается от случая с китайским языком, когда я не понимаю, просто тем, что в первом случае я проделываю больше манипуляций с формальными символами.

Я не показал, что это утверждение ложно, но в данном примере оно определенно должно казаться неправдоподобным. А той правдоподобностью, которой оно все же обладает, оно обязано предположению, будто мы можем построить программу, которая будет иметь те же входы и выходы, что и носители языка, и вдобавок мы исходим из Допущения, что для носителей языка имеется такой уровень описания, на котором они также являются инстанциями программы. На основе данных двух утверждений мы допускаем, что даже если программа Шэнка не исчерпывает всего, чтобы мы могли бы узнать о понимании, она, возможно, есть хотя бы часть этого. Я допускаю, что это эмпирически возможно, но пока никто не привел ни малейшего основания, чтобы могли полагать, что это истинно, мой пример наводит на мысль, — хотя, конечно, и не доказывает, — что компьютерная программа попросту не имеет никакого отношения к моему пониманию рассказа. В случае китайского текста у меня есть все, что может вложить в меня искусственный интеллект посредством программы, но я ничего не понимаю; в случае английского текста я понимаю

все, но пока что нет никаких оснований думать, что мое понимание имеет что-то общее с компьютерными программами, т. е. с компьютерными операциями на элементах, определенных чисто формальным образом. Поскольку программа определена в терминах вычислительных операций на чисто формально определенных элементах, мой пример наводит на мысль, что эти операции сами по себе не имеют интересной связи с пониманием. Они наверняка не составляют достаточных условий, и не было приведено ни малейшего основания считать, что они составляют необходимые условия или хотя бы вносят какой-то существенный вклад в понимание. Заметьте, что сила моего аргумента состоит не просто в том, что различные машины могут иметь одни и те же входы и выходы при том, что их работа основана на разных формальных принципах, — дело вовсе не в этом. Дело в том, что какие бы формальные принципы вы ни закладывали в компьютер, они будут недостаточными для понимания, поскольку человек сможет следовать этим формальным принципам, ничего не понимая. Не было предложено никаких оснований, чтобы думать, будто такие принципы необходимы или хотя бы полезны, поскольку не было дано никаких оснований, чтобы думать, что когда я понимаю английский язык, я вообще оперирую с какой бы то ни было формальной программой.

Но что же все-таки имеется у меня в случае английских предложений, чего у меня нет в случае китайских предложений? Очевидный ответ состоит в том, что в отношении первых я знаю, что они значат, а в отношении вторых у меня нет ни малейшего представления, что они могли бы значить. Но в чем такое представление могло бы состоять и почему мы не могли бы снабдить им машину? Почему машина не могла бы узнать нечто такое обо мне, что означало бы мое понимание английских предложений? Я вернусь к этим вопросам позднее, но сначала хочу продолжить свой пример.

У меня были случаи представить этот пример некоторым людям, работающим в области AI, и, что интересно, они, по-видимому, оказались не согласны друг с другом, что считать правильным ответом на него. У меня скопилось поразительное разнообразие ответов, и ниже я рассмотрю самые распространенные из них (классифицированные в соответствии с их географическим происхождением).

Но сначала я хочу упредить некоторые распространенные недоразумения насчет “понимания”: в ряде из этих дискуссий можно найти множество причудливых толкований слова “понимание”. Мои критики указывают, что есть много различных степеней понимания; что “понимание” — это не простой двухместный предикат; что есть различные виды и уровни понимания, и часто даже закон исключенного третьего невозможно простым образом приложить к утверждениям формы “x понимает y”; что во многих случаях вопрос, понимает ли x y, оказывается вопросом нашего решения, а не простым фактическим вопросом и так далее. На все это я хочу ответить: “Конечно, конечно”. Но все это не имеет никакого отношения к обсуждаемым вопросам. Имеются ясные случаи, в которых можно буквально говорить о понимании, и ясные случаи, в которых о нем нельзя говорить; и мне для моей аргументации в этой статье только и нужны эти два вида случаев [4]. Я понимаю рассказы на английском языке; в меньшей степени я понимаю рассказы по-французски; в еще меньшей степени я понимаю рассказы по-немецки; а по-китайски вообще не понимаю. Что же касается моего автомобиля и моей счетной машинки, то они вообще ничего не понимают; они не по этой части. Мы часто метафорически и аналогически атрибутируем “понимание” и другие когнитивные предикаты автомобилям, счетным машинам и другим артефактам, но такие атрибуции ничего не доказывают. Мы говорим: “Дверь знает, когда открываться, так как в ней есть фотоэлемент”, “Счетная

машинка знает (умеет, способна), как складывать и вычитать, но не делить” и “Термостат воспринимает изменения температуры”. Очень интересно, на каком основании мы делаем такие атрибуции, и это основание связано с тем, что мы распространяем на артефакты нашу собственную интенциональность *s*; наши инструменты суть продолжения наших целей, и поэтому мы находим естественным

приписывать им метафорическим образом интенциональность [5]; но я считаю, что такие примеры не решают никаких философских вопросов. Тот смысл, в каком автоматическая дверь “понимает инструкции” посредством своего фотоэлемента, — это вовсе не тот смысл, в каком я понимаю английский язык. Если имеется в виду, что программированный компьютер Шэнка понимает рассказы в том же метафорическом смысле, в каком понимает дверь, а не в том смысле, в каком я понимаю английский язык, то этот вопрос не стоит и обсуждать. Но Ньюэлл и Саймон (1963) пишут, что познание, которое они атрибутируют машинам, есть в точности то познание, которое присуще людям. Мне нравится прямота этой претензии, и именно эту претензию я буду рассматривать. Я буду аргументировать, что в буквальном смысле слова программированный компьютер понимает ровно столько, сколько автомобиль и счетная машинка, то есть ровным счетом ничего. Понимание чего бы то ни было компьютером не просто частично или неполно (подобно моему пониманию немецкого языка); оно попросту равно нулю.

А теперь рассмотрим ответы:

I. ОТВЕТ ОТ СИСТЕМ (БЕРКЛИ)

“Это правда, что тот человек, который заперт в комнате, не понимает рассказа, но дело в том, что он всего лишь часть некоей цельной системы, и эта система понимает рассказ. Перед ним лежит грассбух, в котором записаны правила, у него имеется стопка бумаги и чернила, чтобы делать вычисления, у него есть “банки данных” — наборы китайских символов. И вот понимание приписывается не просто индивиду; оно приписывается всей этой системе, частью которой он является”.

Мой ответ теории систем очень прост: позвольте вашему индивиду интериоризовать все эти элементы системы. Пусть он выучит наизусть все правила из грассбуха и все банки данных — все китайские символы, и пусть делает вычисления в уме. Тогда индивид будет воплощать в себе всю систему. Во всей системе не останется ничего, что он не охватил бы в себе. Мы можем даже отбросить комнату и допустить, что он работает под открытым небом. Все равно он абсолютно не понимает китайский язык, и тем более этот язык не понимает система, ибо в системе нет ничего, чего не было бы в нем. Если он не понимает, то система никаким образом не сможет понимать, ибо эта система — всего лишь часть его. На самом деле, мне неловко давать даже такой ответ представителям теории систем, ибо эта теория кажется мне слишком неправдоподобной, чтобы начинать с нее. Ее идея состоит в том, что если некий человек не понимает китайского языка, то каким-то образом объединение этого человека с листками бумаги могло бы понимать китайский язык. Мне нелегко вообразить себе, как вообще человек, не зашоренный некоей идеологией, может находить эту идею правдоподобной. И все же я думаю, что многие люди, связавшие себя с идеологией сильного AI, склонны будут в конечном счете сказать нечто очень похожее на это; поэтому давайте рассмотрим эту идею еще чуть-чуть подробнее. Согласно одному из вариантов данного взгляда, если человек из примера с интериоризованной системой и не понимает китайского языка в том смысле, в каком его понимает носитель китайского языка

(потому что, например, он не знает, что в этом рассказе упоминаются рестораны и гамбургеры и т. д.), все же “этот человек как система манипулирования формальными символами” на самом деле понимает китайский язык. Ту подсистему этого человека, которая ответственна за манипуляцию с формальными символами китайского языка, не следует смешивать с его подсистемой английского языка.

Таким образом, в этом человеке на самом деле оказываются две подсистемы: одна понимает английский язык, а другая — китайский, и “все дело в том, что эти две системы мало как связаны друг с другом”. Но я хочу ответить, что не только они мало связаны друг с другом, но между ними нет даже отдаленного сходства. Та подсистема, которая понимает английский язык (допустим, что мы позволили себе на минуту разговаривать на этом жаргоне “подсистем”) знает, что наши рассказы суть о ресторанах и поедании гамбургеров, она знает, что ей задают вопросы о ресторанах и что она отвечает на эти вопросы, используя все свои возможности, делая различные выводы из содержания рассказа, и так далее. Но китайская система ничего этого не знает. Тогда как английская система знает, что “hamburgers” указывает на гамбургеры, китайская подсистема знает лишь, что за такой-то загогулиной следует такая-то закорючка. Она знает только, что на одном конце вводятся различные формальные символы и ими манипулируют по правилам, записанным на английском языке, а на другом конце выходят другие символы. Весь смысл нашего исходного примера состоял в том, чтобы аргументировать, что такой манипуляции символами самой по себе недостаточно для понимания китайского языка в каком бы то ни было буквальном смысле, потому что человек может рисовать такую-то закорючку вслед за такой-то загогулиной, ничего не понимая по-китайски. И постулирование подсистем в человеке не составляет хорошего контраргумента, потому что подсистемы Для нас вовсе не лучше самого человека; они по-прежнему не обладают ничем таким, что хотя бы отдаленно напоминало то, чем обладает говорящий по-английски человек (или подсистема). В сущности, в описанном нами случае китайская подсистема есть попросту часть английской подсистемы — та часть, которая манипулирует бессмысленными символами в соответствии с правилами, записанными на английском языке.

Спросим себя, какова основная мотивация ответа от теории систем; какие независимые основания имеются, как предполагается, дабы утверждать, что агент должен иметь в себе некую подсистему, которая понимает (в буквальном смысле слова “понимать”) рассказы на китайском языке? Насколько я могу судить, единственное основание состоит в том, что в нашем примере у меня имеются те же самые вход и выход, что у настоящих носителей китайского языка, и программа, приводящая от входа к выходу. Но вся суть наших примеров состояла в том, чтобы попытаться показать, что этого недостаточно для понимания — в том смысле слова “понимание”, в каком я понимаю рассказы на английском языке, ибо человек, а стало быть, и набор систем, вместе составляющих человека, могут обладать правильной комбинацией входа, выхода и программы и все же ничего не понимать — в том относящемся к делу буквальном смысле слова “понимать”, в каком я понимаю английский язык. Единственная мотивация утверждения, что во мне должна быть некая подсистема, понимающая китайский язык, состоит в том, что я имею некую программу и я успешно прохожу тест Тьюринга; я могу дурачить настоящих носителей китайского языка. Но мы как раз и обсуждаем, среди прочего, адекватность теста Тьюринга. Наш пример показывает, что может случиться так, что есть две “системы” — обе успешно проходят тест Тьюринга, но лишь одна из них действительно понимает; и никудышным контраргументом было бы сказать, что раз обе успешно проходят тест Тьюринга, обе должны понимать, поскольку это утверждение не согласуется с аргументом,

гласящим, что та система во мне, которая понимает английский язык, обладает чем-то гораздо большим, чем та система, которая просто оперирует с китайским текстом. Короче говоря, ответ от теории систем попросту уклоняется от сути спора, неаргументированно настаивая на том, что данная система должна понимать по-китайски.

Кроме того, ответ от теории систем, по-видимому, ведет к таким последствиям, которые абсурдны по независимым от нашего спора основаниям. Если мы собираемся сделать вывод, что во мне имеется некое познание, на том основании, что у меня имеются вход и выход, а между ними — программа, то тогда, видимо, окажется, что все и всяческие некогнитивные подсистемы станут когнитивными. Например, на некотором уровне описания мой желудок занимается обработкой информации, и он инстанцирует сколько угодно компьютерных программ, но я так понимаю, что мы не хотели бы сказать, что мой желудок что-то понимает (ср. Pylyshyn 1980). Но если мы примем ответ от теории систем, то трудно видеть, как нам избежать утверждения, будто желудок, сердце, печень и т. д. суть понимающие подсистемы, ибо нет никакого принципиального способа отличать мотивацию утверждения, что китайская подсистема обладает пониманием, от утверждения, что желудок обладает пониманием. Не будет, кстати, хорошим ответом на это, если мы скажем, что китайская система имеет на входе и выходе информацию, а желудок имеет на входе пищу и на выходе продукты пищеварения, ибо с точки зрения агента, с моей точки зрения, информации нет ни в пище, ни в китайском тексте, китайский текст — это попросту скопище бессмысленных закорючек. В случае китайского примера информация имеется только в глазах программистов и интерпретаторов, и ничто не может помешать им толковать вход и выход моих пищеварительных органов как информацию, если они этого пожелают.

Это последнее соображение связано с некоторыми независимыми проблемами в сильном AI и имеет смысл отвлечься на минуту, чтобы разъяснить их. Если сильный AI хочет быть отраслью психологии, он должен уметь отличать действительно ментальные системы от тех, что ментальными не являются. Он должен отличать принципы, на которых основывается работа сознания, от принципов, на которых основывается работа нементальных систем; в противном случае он не даст нам никаких объяснений того, что же такого специфически ментального в ментальном. И дистинкция “ментальное — нементальное” не может зависеть только от точки зрения внешнего наблюдателя — она должна быть внутренне присущей самим системам; иначе любой наблюдатель имел бы право, если бы пожелал, трактовать людей как нементальные феномены, а, скажем, ураганы — как ментальные. Но очень часто в литературе по AI эта дистинкция смазывается таким образом, что в конечном счете это смазывание может оказаться фатальным для претензий AI быть когнитивным исследованием. Маккарти, например, пишет: “Можно сказать, что такие простые машины, как термостаты, обладают убеждениями (beliefs), и обладание убеждениями присуще, кажется, большинству машин, способных решать задачи” (McCarthy 1979). Всякий, кто считает, что сильный AI имеет шанс стать теорией сознания, должен поразмыслить над импликациями этого замечания. Нас просят принять в качестве открытия, сделанного сильным AI, что кусок металла на стене, употребляемый нами для регулирования температуры, обладает убеждениями в точности в том же самом смысле, в каком мы, наши супруги и наши дети обладают убеждениями, и более того — что “большинство” других машин в комнате: телефон, магнитофон, калькулятор, выключатель лампочки — также обладают убеждениями в этом буквальном смысле. В цели данной

статьи не входит аргументировать против замечания Маккарти, так что я просто без аргументации выскажу следующее.

Исследование сознания начинается с таких фактов, как то, что люди обладают убеждениями, а термостаты, телефоны и счетные машинки не обладают. Если вы получаете теорию, отрицающую этот факт, то это означает, что вы построили контрпример данной теории, и она ложна. Создается впечатление, что те люди, работающие в AI, которые пишут такие вещи, думают, что могут легко отбросить их, ибо на самом деле не принимают их всерьез, и они не думают, что другие примут их всерьез. Я же предлагаю принять их всерьез — хотя бы на минуту. Поразмыслите напряженно в течение одной минуты, что именно необходимо, дабы установить, что вот этот кусок металла, висящий на стене, обладает настоящими убеждениями: убеждениями с направлением соответствия, пропозициональным содержанием и условиями выполнимости; убеждениями, могущими быть сильными или слабыми; нервными, тревожными или безмятежными убеждениями; догматическими, рациональными или суеверными убеждениями; слепой верой или сомневающимся познанием; убеждениями какого угодно рода. Термостат — не кандидат на обладание такими убеждениями. Равно как и желудок, печень, счетная машинка или телефон. Однако раз мы принимаем эту идею всерьез, заметьте, что истинность его была бы фатальной для претензий сильного AI быть наукой о сознании. Ибо теперь сознание — повсюду. Мы-то хотели узнать, чем отличается сознание от термостатов и печеней. И если бы Маккарти оказался прав, сильный AI не имел бы надежды сообщить нам это.

2. ОТВЕТ ОТ РОБОТА (ЙЕЙЛ)

“Предположим, мы написали программу, отличную от программы Шэнка. Предположим, мы поместили компьютер внутрь некоего робота, и этот компьютер не просто воспринимал бы формальные символы на входе и выдавая бы формальные символы на выходе, а на самом деле руководил бы роботом так, что тот делал бы нечто очень похожее на сенсорное восприятие, хождение, движение, забивание гвоздей, еду и питье — в общем, все что угодно. Этот робот, к примеру, имел бы встроенную телекамеру, которая давала бы ему возможность "видеть", он имел бы руки и ноги, которые давали бы ему возможность "действовать", и все это управлялось бы его компьютерным "мозгом". Такой робот, в отличие от компьютера Шэнка, обладал бы настоящим пониманием и другими ментальными состояниями”.

Первое, на что следует обратить внимание в ответе от робота, вот что: этот ответ молчаливо соглашается с тем, что понимание — вопрос не только манипуляций с формальными символами, ибо этот ответ добавляет некий набор причинных отношений с внешним миром. Но ответ на ответ от робота заключается в том, что добавление таких “перцептуальных” или “моторных” способностей ничего не добавляет к исходной программе Шэнка в том, что касается понимания в частности или интенциональности вообще. Чтобы увидеть это, обратите внимание на то, что тот же самый мысленный эксперимент приложим и к случаю с роботом. Предположим, что вместо того чтобы помещать компьютер внутрь робота, вы помещаете меня внутрь комнаты, и — как и в первоначальном случае с китайскими текстами — вы даете мне еще больше китайских символов с еще большим количеством инструкций на английском языке насчет того, как сопоставлять одни китайские символы с другими и выдавать китайские символы вовне. Предположим далее, что некоторые китайские символы, приходящие ко мне от телекамеры, встроенной в робота, и другие китайские символы, которые выдаю я, служат

для того, чтобы включать моторы, встроенные в робота, так чтобы двигались ноги и руки робота, но я ничего этого не знаю. Важно подчеркнуть, что я делаю только одно — манипулирую формальными символами: я не знаю никаких дополнительных фактов. Я получаю “информацию” от “перцептивного” аппарата робота и выдаю “инструкции” его моторному аппарату, ничего этого не зная. Я — гомункулус этого робота, но в отличие от традиционного гомункулуса, я не знаю, что происходит. Я не понимаю ничего, кроме правил манипулирования символами. И вот в этом случае я хочу сказать, что у нашего робота нет никаких интенциональных состояний; он двигается просто в результате функционирования своей электросхемы и ее программы. И более того, инстанцируя эту программу, я не имею никаких интенциональных состояний соответствующего рода. Я только следую инструкциям о манипулировании формальными символами.

3. ОТВЕТ ОТ МОДЕЛИРОВАНИЯ РАБОТЫ МОЗГА (БЕРКЛИ И МАССАЧУСЕТССКИЙ ТЕХНОЛОГИЧЕСКИЙ ИНСТИТУТ)

“Предположим, мы построили программу, которая не репрезентирует информацию, имеющуюся у нас о мире, — вроде той информации, которая имеется в сценариях Шэнка, но которая моделирует действительную последовательность возбуждения нейронов в синапсах мозга настоящего носителя китайского языка, когда он понимает рассказы на китайском языке и дает ответы на них. Машина принимает на входе китайские рассказы и вопросы к ним, моделирует структуру настоящего китайского мозга, когда он информационно обрабатывает эти рассказы, и выдает на выходе китайские ответы. Мы можем даже вообразить, что эта машина оперирует не одной-единственной последовательно действующей программой, а целым набором параллельно действующих программ — подобно тому, как, видимо, функционирует настоящий мозг человека, когда он информационно обрабатывает естественный язык. И вот в этом случае мы

наверняка должны были бы сказать, что наша машина понимает рассказы; а если мы откажемся признать это, то не должны ли мы будем также отрицать, что настоящие носители китайского языка понимают эти рассказы? На уровне синапсов чем отличались бы или чем могли бы отличаться программа компьютера и программа китайского мозга?”

Прежде чем излагать свои возражения на этот ответ, я хотел бы отвлечься и заметить, что этот ответ странен в устах сторонника искусственного интеллекта (или функционализма и т. д.): я-то думал, что вся идея сильного AI заключается в том, что для того чтобы знать, как работает сознание, нам не нужно знать, как работает мозг. Базовая гипотеза состояла в том (или, по крайней мере, я предполагал, что она состоит в том), что имеется некий уровень ментальных операций, состоящих из вычислительных процессов над формальными элементами, и этот уровень представляет собой сущность ментального, и он может быть реализован в самых разнообразных мозговых процессах — точно так же, как любая компьютерная программа может быть реализована с помощью самых разных аппаратных устройств: согласно допущениям сильного AI, сознание относится к телу как программа относится к аппаратному устройству компьютера, и стало быть, мы в состоянии понимать сознание, не занимаясь нейрофизиологией. Если бы для того чтобы заниматься искусственным интеллектом, нам нужно было бы предварительно знать, как работает мозг, то нам не к чему было бы и утруждать себя занятиями искусственным интеллектом. Однако даже если мы подойдем столь близко к работе мозга, этого не будет достаточно, чтобы продуцировать понимание. Дабы увидеть это, вообразите, что вместо одноязычного человека, сидящего в комнате и тасующего символы, у нас есть человек, который

оперирует неким сложным набором шлангов, связанных друг с другом с помощью клапанов. Когда этот человек получает китайские символы, он сверяется в программе, написанной на английском языке, какие клапаны ему нужно открыть и какие закрыть. Каждое соединение шлангов соответствует некоему синапсу в китайском мозгу, и вся система устроена так, что после возбуждений всех нужных синапсов, т. е. после отворачивания всех нужных кранов, на выходе всей этой последовательности шлангов выскакивают китайские ответы.

Ну и где же в этой системе понимание? Она принимает на входе китайские тексты, моделирует формальную структуру синапсов китайского мозга и выдает китайские ответы на выходе. Но этот наш человек определенно не понимает китайского языка, и эти наши шланги не понимают китайского языка, и если мы соблазнили бы согласиться с взглядом, который я считаю нелепым, — именно взглядом, согласно которому каким-то образом обладает пониманием конъюнкция этого человека и шлангов, то вспомните, что в принципе человек может интериоризовать формальную структуру шлангов и проделывать все эти “возбуждения нейронов” в своем воображении. Проблема с моделированием работы мозга состоит в том, что моделируется не то в работе мозга, что нужно. Поскольку моделируется только формальная структура последовательности нейронных возбуждений в синапсах, то не моделируется та сторона функционирования мозга, которая как раз и имеет значение, именно: каузальные свойства мозга, его способность продуцировать интенциональные состояния. А то, что формальных свойств недостаточно для каузальных свойств, показывает пример с шлангами: здесь все формальные свойства налицо, но они отсечены от интересных для нас нейробиологических каузальных свойств.

4. КОМБИНИРОВАННЫЙ ОТВЕТ (БЕРКЛИ И СТЭНФОРД)

“Пусть каждый из трех предыдущих ответов, возможно, не является абсолютно убедительным в качестве опровержения контрпримера с китайской комнатой; но если вы объедините все три ответа, то вместе они гораздо более убедительны и даже бесспорны. Вообразите робота, в черепной полости которого встроен имеющий форму мозга компьютер; вообразите, что этот компьютер запрограммирован всеми синапсами человеческого мозга; вообразите далее, что все поведение робота неотлично от человеческого поведения; и вот взгляните на всю эту штуку как на единую систему, а не просто как на компьютер с входами и выходами. Наверняка в этом случае мы должны были бы приписать нашей системе ин-тенциональность”.

Я полностью согласен, что в таком случае мы сочли бы, что рационально и даже неизбежно принятие гипотезы о том, что данный робот обладает интенциональностью, поскольку мы ничего больше не знали бы о нем. В самом деле, никакие иные элементы этой комбинации, кроме внешнего вида и поведения, не имеют значения. Если бы нам удалось построить робота, поведение которого было бы на протяжении крупного интервала времени неотлично от человеческого поведения, мы атрибутировали бы ему интенциональность — до тех пор, пока не получили бы каких-либо оснований не делать этого. Нам не нужно было бы заранее знать, что его компьютерный мозг — формальный аналог человеческого мозга.

Но я действительно не вижу, как это могло бы помочь претензиям сильного AI, и вот почему. Согласно сильному AI, инстанцирование формальной программы с подходящими входом и выходом представляет собой достаточное условие

интенциональности — и в сущности, конституирует интенциональность. Как говорит Ньюэлл (Newell

1979), "сущность ментального заключается в оперировании некоей физической системой символов. Но то атрибутивное интенционально-сти роботу, которое мы совершаем в этом примере, не имеет ничего общего с формальными программами. Оно основывается просто на допущении, что если робот выглядит и действует в достаточной степени подобно нам, то мы предположили бы, — пока не доказано обратное, — что он должен иметь ментальные состояния, подобные нашим, причиняющие его поведение и выражаемые в этом поведении, и он должен иметь некий внутренний механизм, способный продуцировать такие ментальные состояния. Если бы мы независимым образом знали, как объяснить его поведение без таких допущений, мы бы не атрибутировали ему интенциональность, в особенности, если бы мы знали, что у него имеется формальная программа. А это как раз и есть сущность моего изложенного выше ответа на возражение 2.

Предположим, что мы знаем, что поведение робота полностью объясняется тем фактом, что некий человек, сидящий внутри него, получает неинтерпретированные формальные символы от сенсорных рецепторов робота и посылает неинтерпретированные формальные символы его двигательным механизмам, и человек этот совершает свои манипуляции с символами в соответствии с неким набором правил. Предположим далее, что данный человек не знает всех этих фактов о роботе; все, что он знает, — какие операции нужно совершать над тем или иным бессмысленным символом. В таком случае мы бы сочли робота хитроумно сконструированной механической куклой. Гипотеза о том, что эта кукла обладает сознанием, была бы теперь ничем не подкрепленной и излишней, ибо нет никаких оснований приписывать интенциональность этому роботу или системе, частью которой он является (кроме, разумеется, интенциональности человека при его манипулировании символами). Манипуляции с формальными символами продолжаются, выход правильным образом соответствует входу, но единственный действительный локус интенциональности — человек, сидящий в кукле, а он не знает ни одного из относящихся к делу интенциональных состояний; он, к примеру, не видит, что отражается в глазах робота, он не намеревается двигать рукой робота, и он не понимает ни одного из замечаний, которые делает робот или которые ему адресуются. И, по изложенным выше основаниям, ничего этого не делает та система, частями которой являются человек и робот.

Чтобы понять это соображение, сопоставьте этот случай со случаями, в которых мы находим совершенно естественным приписывать интенциональность членам некоторых других видов приматов — например, обезьянам и таким домашним животным, как собаки. Оснований, по которым мы находим это естественным, грубо говоря, две : мы не можем понять поведение этих животных без приписывания им

интенциональности, и мы видим, что эти звери сделаны из материала, похожего на тот, из которого сделаны мы сами: это - глаз, это — нос, а вот это — его кожа и так далее. При том, что поведение животного обладает связностью и последовательностью, а также при допущении, что в его основе лежит тот же самый каузальный материал, мы допускаем, что в основе поведения этого животного должны быть ментальные состояния, и эти ментальные состояния должны продуцироваться механизмами, сделанными из материала, подобного нашему материалу. Мы наверняка сделали бы такие же допущения о роботе, если бы у нас не было оснований не делать их, но раз мы узнали, что его поведение есть результат формальной программы и действительные каузальные свойства физического

вещества не имеют к этому отношения, мы отбросили бы допущение об интенциональности. (См, Multiple authors 1978.)

Часто встречаются еще два ответа на мой пример (и поэтому заслуживают обсуждения), но на самом деле и они бьют мимо цели.

5. ОТВЕТ ОТ ДРУГИХ СОЗНАНИЙ (ЙЕЙЛ)

“Откуда вы знаете, что другие люди понимают китайский язык или что-то еще? Только по их поведению. И вот компьютер может пройти поведенческие тесты столь же успешно, как и люди (в принципе), так что если вы собираетесь приписывать познание другим людям, вы должны в принципе атрибутировать его и компьютерам”.

Это возражение заслуживает лишь короткой реплики. Обсуждаемая нами проблема состоит не в том, откуда я знаю, что у других людей имеются когнитивные состояния, а в том, что именно я им атрибутирую, когда атрибутирую когнитивные состояния. Суть моего аргумента в том, что это не может быть всего лишь вычислительными процессами и их выходом, ибо вычислительные процессы и их выход могут существовать без когнитивного состояния. Притворяться бесчувственным — это не ответ на данный аргумент. В “когнитивных науках” предполагается реальность и познаваемость ментального точно так же, как в физических науках должно предполагать реальность и познаваемость физических объектов.

6. ОТВЕТ ОТ НЕСКОЛЬКИХ ОБИТАЛИЦ (БЕРКЛИ)

“Весь ваш аргумент исходит из предпосылки, что AI ведет речь только об аналоговых и цифровых компьютерах. Но это всего лишь современное состояние технологии. Что бы ни представляли собой те каузальные процессы, которые, как вы говорите, существенны для интенциональности (при допущении, что вы правы), в конечном счете мы научимся создавать устройства, обладающие этими каузальными процессами, и это будет искусственным интеллектом. Так что ваши аргументы никоим образом не направлены на способность искусственного интеллекта продуцировать и объяснять познание”.

На самом деле у меня нет возражений против этого ответа, разве что одно: он на самом деле тривиализует замысел сильного AI, переопределяя его как все то, что искусственно продуцирует и объясняет познание. Важность же первоначальной претензии, заявленной от лица искусственного интеллекта, состояла в том, что она представляла собой точный вполне определенный тезис: ментальные процессы суть вычислительные процессы над формально определенными элементами. Моя цель заключалась в том, чтобы оспорить этот тезис. Если же претензия переопределена так, что она более не совпадает с этим тезисом, то и мои возражения более неприложимы, потому что у нас в руках не остается никакой проверяемой гипотезы, к которой их можно было бы приложить.

Вернемся теперь к вопросу, на который я обещал попробовать ответить: в предположении, что в моем самом первом примере я понимаю английский язык и не понимаю китайский, и поэтому машина не понимает ни по-английски, ни по-китайски, все же должно быть во мне нечто, благодаря чему истинно, что я понимаю английский язык, и также должно быть во мне соответствующее нечто, благодаря чему истинно, что не понимаю китайский. И почему бы мы не могли передать эти нечто, — чем бы они ни были, — машине?

Я в принципе не вижу никаких оснований, почему бы мы не могли передать машине способность понимать английский или китайский языки, ибо в некотором важном смысле наши тела и наши мозги суть в точности такие же самые машины. Но с другой стороны я вижу очень сильные основания для утверждения, что мы не смогли бы передать такие способности машине в том случае, когда функционирование этой машины определено только в терминах вычислительных процессов над формально определенными элементами; то есть когда функционирование машины определено как инстанциация некоей компьютерной программы. Не потому я способен понимать английский язык и имею еще другие формы интенциональности, что я являюсь инстанциацией компьютерной программы (я, наверное, являюсь инстанциацией какого угодно числа компьютерных программ), но — насколько мы знаем — потому, что я являюсь организмом определенного рода с определенной биологической (т. е. химической и физической) структурой, и эта структура, при определенных условиях, каузально способна продуцировать восприятие, действие, понимание, обучение и другие интенциональные феномены. И один аспект предлагаемого мной аргумента состоит в том, что лишь нечто такое, что обладает этими каузальными способностями и, могло бы обладать интенциональностью. Быть может, другие физические и химические процессы могли бы продуцировать в точности те же эффекты; быть может, к примеру, марсиане также обладают интенциональностью, но их мозги сделаны из иного вещества.

Это — эмпирический вопрос, подобно вопросу о том, может ли фотосинтез осуществляться чем-то имеющим химическое строение, отличное от химического строения хлорофилла.

Но основной пункт предлагаемого мной аргумента состоит в том, что никакая чисто формальная модель никогда сама по себе не будет достаточна для интенциональности, потому что формальные свойства сами по себе не конституируют интенциональность, и они сами по себе не обладают каузальными способностями, за исключением способности порождать, будучи инстанцированными, следующую стадию формализма, когда машина запущена и работает. И любые другие каузальные способности, имеющиеся у конкретных реализаций формальной модели, не имеют отношения к самой этой формальной модели, потому что мы всегда можем поместить эту же самую формальную модель в какую-нибудь иную реализацию, в которой эти каузальные свойства очевидным образом отсутствуют. Даже если носители китайского языка каким-то чудом в точности реализуют программу Шэнка, мы можем поместить ту же самую программу в носителей английского языка, в шланги или в компьютеры — а ведь ни то, ни другое, ни третье не понимает китайского языка, несмотря на программу.

В функционировании мозга важна не та формальная тень, которую отбрасывает последовательность синапсов, а действительные свойства этих последовательностей. Все знакомые мне аргументы в пользу сильного варианта AI настаивают на том, чтобы нарисовать некий абрис, следуя тени, отбрасываемой познанием, и затем утверждать, что эти тени и суть та самая штука, тенями которой они являются.

В заключение я хочу попытаться сформулировать некоторые общефилософские соображения, неявно присутствующие в моем аргументе. Для ясности я постараюсь построить изложение в виде вопросов и ответов и начну с одного избитого вопроса, а именно:

“Может ли машина мыслить?”

Ответ очевидным образом положителен. Мы как раз и есть такие машины.

“Да, но может ли мыслить артефакт, машина, сделанная человеком?” В предположении, что возможно искусственно произвести машину с нервной системой, нейронами, обладающими аксонами и дендритами, и со всем прочим, в достаточной степени похожими на нашу нервную систему, наши нейроны и т. д., ответ на этот вопрос представляется опять же тривиально положительным. Если вы в состоянии точно продублировать причины, то вы в состоянии продублировать и следствия. И на самом деле, возможно, быть может, продуцировать сознание, интенциональность и все такое прочее, используя какие-то другие химические принципы, чем те, что реализованы в людях. Это, как я сказал, вопрос эмпирический.

“Хорошо, ну а может ли мыслить цифровой компьютер?”

Если под “цифровым компьютером” мы понимаем любой предмет, имеющий такой уровень описания, на котором его можно корректно описать как инстанциацию компьютерной программы, то ответ, разумеется, опять же положителен, ибо мы суть инстанциации какого угодно числа компьютерных программ и мы можем мыслить.

“Но может ли какой-нибудь предмет мыслить, понимать и так далее только в силу того, что этот предмет — компьютер с подходящей программой? Может ли свойство инстанцирования программы — подходящей программы, конечно, — быть само по себе достаточным условием понимания?”

Вот это, по-моему, хороший вопрос, хотя обычно его путают с каким-нибудь из тех вопросов, что приведены выше, и ответ на него отрицателен.

“Но почему?”

Потому что манипуляция формальными символами сама по себе не обладает никакой интенциональностью; она лишена смысла; она даже не является манипуляцией символами, ибо эти символы ничего не символизируют. Используя лингвистический жаргон, можно сказать, что они имеют лишь синтаксис, но не имеют семантики. Такая интенциональность, которой, как кажется, обладает компьютер, существует единственно в мозгах тех людей, которые запрограммировали его и используют его, посылают нечто на вход и интерпретируют то, что появляется на выходе.

Цель примера с китайской комнатой состояла в том, чтобы попытаться показать это, показав, что как только мы помещаем нечто в систему, которая на самом деле обладает интенциональностью (человек), и программируем эту систему формальной программой, то вы видите, что эта формальная программа не несет никакой дополнительной интенциональности. Она ничего не прибавляет, например, к способности человека понимать китайский язык.

Именно эта черта AI — различие между программой и реализацией, казавшаяся столь привлекательной, — оказывается фатальной для претензии на то, что моделирование может стать дублированием. Различие между программой и ее реализацией в аппаратном устройстве компьютера, по-видимому, параллельно различию между уровнем ментальных

операций и уровнем мозговых операций. И если бы мы могли описать уровень ментальных операций как формальную программу, то, представляется, мы могли бы описать суть сознания, не занимаясь интроспективной психологией или нейрофизиологией мозга. Но уравнение “сознание относится к мозгу так же, как программа к аппаратному устройству компьютера” рушится в нескольких пунктах, в том числе в следующих трех: во-первых, из различия между программой и ее реализацией следует, что одна и та же программа могла бы обладать самыми сумасшедшими реализациями, которым неприсуща форма интенциональности. Визенбаум (Wiezenbaum 1976: Ch. 2), например, подробно рассказывает, как построить компьютер, используя рулон туалетной бумаги и горсть камешков. Точно так же программу, понимающую китайские рассказы, можно запрограммировать в последовательность шлангов, в множество ветряных мельниц или в человека, говорящего только на английском языке, — и тогда ни первая, ни второе, ни третий не станут понимать китайский язык. Начать с того, что камни, туалетная бумага, ветер и шланги — не те вещи, что могут быть наделены интенциональностью (ею могут быть наделены лишь предметы, обладающие теми же каузальными способностями, что и мозг), и хотя носитель английского языка сделан из подходящего для интенциональности материала, нетрудно видеть, что он не приобретает никакой дополнительной интенциональности, выучив наизусть программу, поскольку выучивание ее наизусть не научит его китайскому языку;

во-вторых, программа чисто формальна, а интенциональные состояния неформальны в этом смысле. Они определены в терминах их содержания, а не формы. Убеждение, что идет дождь, например, определено не как некая форма, а как определенное ментальное содержание с условиями выполнения, направлением соответствия (ср. Searle 1979a) и тому подобным. В сущности, убеждение как таковое даже и не имеет формы в этом синтаксическом смысле, ибо одному и тому же убеждению можно придать неопределенно большое число различных синтаксических выражений в различных языковых системах;

в-третьих, как я отметил выше, ментальные состояния и события суть в буквальном смысле продукты функционирования мозга, но программа не есть продукт работы компьютера в этом же смысле.

“Хорошо, но если программы никоим образом не конституируют ментальные процессы, то почему столь многие думали наоборот? Это ведь нужно хоть как-то объяснить”.

На самом деле я не знаю ответа на этот вопрос. Идея, что компьютерные модели могут быть самими реальными вещами, должна была бы показаться подозрительной — прежде всего потому, что компьютер, во всяком случае, не ограничивается моделированием ментальных операций. Никому ведь не приходит в голову, что компьютерное моделирование пожарной тревоги может сжечь дотла соседние дома или что компьютерное моделирование ливня заставит нас всех промокнуть. Так почему же кому-то должно прийти в голову, что компьютерная модель понимания на самом деле что-то понимает? Иногда говорят, что заставить компьютер почувствовать боль или влюбиться — ужасно трудная задача, но любовь и боль ни труднее, ни легче, чем познание или что-то еще. Все, что вам нужно для моделирования — это подходящий вход, подходящий выход и между ними подходящая программа, преобразующая первое во второе. Что бы ни делал компьютер, ничего, кроме этого, у него нет. Спутать моделирование чего-то с

дублированием этого самого — ошибка одного и того же рода, идет ли речь о моделях боли, любви, познания, пожара или ливня.

И все же есть несколько оснований, почему должно было казаться, — а многим людям, возможно, и сейчас кажется, — что AI каким-то образом воспроизводит и тем самым объясняет ментальные феномены, и я полагаю, что нам не удастся устранить эти иллюзии, пока мы полностью не выявим основания, их порождающие.

Первое (и, быть может, самое важное) основание — это путаница с понятием “обработка информации”: многие, занимающиеся когнитивной наукой, полагают, что мозг человека с его сознанием занят чем-то таким, что называется “обработкой информации”, и точно так же компьютер со своей программой занят обработкой информации; пожары же и ливни, с другой стороны, не занимаются обработкой информации. Таким образом, хотя компьютер может моделировать формальные стороны какого угодно процесса, он стоит в некоем особом отношении к сознанию и мозгу, ибо когда компьютер подходящим образом запрограммирован в идеале той же программой, что и мозг, то обработка информации тождественна в обоих случаях, и такая обработка информации и есть на самом деле сущность ментального. Но с этим аргументом беда в том, что он основывается на двусмысленности понятия “информация”. В том смысле, в каком люди “обрабатывают информацию”, когда они размышляют, скажем, над арифметическими

задачками или когда они читают рассказы и отвечают на вопросы о них, — в этом смысле запрограммированный компьютер вовсе не занимается никакой “обработкой информации”. Вместо этого он манипулирует формальными символами. Тот факт, что программист и интерпретатор компьютерного выхода используют символы для замещения неких объектов в мире, — не имеет никакого отношения к самому компьютеру. Компьютер, повторим, имеет синтаксис, но лишен семантики. Так, если вы напечатаете компьютеру: “Сколько будет 2×2 ?”, то он вам напечатает “4”. Но он не имеет никакого представления о том, что “4” означает 4 или что “4” вообще означает что бы то ни было. И дело не в том, что ему не хватает какой-нибудь второпорядковой информации об интерпретации его первопорядковых символов, а в том, что его первопорядковые символы лишены всякой интерпретации, пока речь идет о компьютере. Все, что имеется у компьютера, — это символы и еще раз символы. Поэтому введение понятия “обработка информации” ставит нас перед дилеммой: либо мы толкуем понятие “обработка информации” таким образом, что оно влечет интенциональность как часть процесса обработки, либо мы его так не толкуем. Если первое, то запрограммированный компьютер не занимается обработкой информации, а лишь манипулирует формальными символами. Если второе, то хотя компьютер занимается обработкой информации, он совершает ее лишь в том смысле, в каком ее совершают счетные машинки, термостаты, ливни и ураганы; именно у всех у них имеется такой уровень описания, на котором мы можем описать их так, что они принимают информацию на одном конце, преобразуют ее и продуцируют информацию на выходе. Но в этом случае интерпретировать их вход и выход как информацию в обычном смысле этого слова приходится внешним наблюдателям. И тогда в терминах сходства процессов обработки информации не удастся установить сходство между компьютером и мозгом.

Во-вторых, в большей части исследований по AI наличествуют остатки позиций бихевиоризма или функционализма. Поскольку подходящим образом запрограммированные компьютеры могут иметь схемы входа и выхода, сходные со схемами входа и выхода у людей, у нас появляется соблазн постулировать у компьютеров ментальные состояния,

сходные с человеческими ментальными состояниями. Но раз мы видим, что возможно и концептуально и эмпирически, чтобы система обладала человеческими способностями в некоей области, вовсе не обладая при этом интенциональностью, то мы должны суметь превозмочь этот соблазн. Мой настольный калькулятор обладает способностями к счету, но не обладает интенциональностью, и в этой статье я попытался показать, что система может быть способной иметь такие вход и выход, которые дублируют вход и выход настоящего носителя

китайского языка, — и все же не понимать по-китайски, как бы она ни была программирована. Тест Тьюринга типичен для этой традиции тем, что он бессовестно бихевиористичен и операционалистичен, и я полагаю, что если бы исследователи AI полностью отреклись от бихевиоризма и операционализма, то исчезла бы большая часть путаницы насчет моделирования и дублирования.

В-третьих, этот остаточный операционализм соединяется с остаточной формой дуализма; в сущности, сильный AI имеет смысл лишь в том случае, если принимается дуалистическое допущение о том, что там, где дело идет о сознании, мозг не имеет значения. В сильном AI (а также и в функционализме) значение имеют программы, а программы независимы от их реализаций в машинах; в сущности, поскольку речь идет об AI, одна и та же программа могла бы быть реализована электронной машиной, картезианской ментальной субстанцией или гегельянским мировым духом. Самое удивительное открытие, которые я сделал, обсуждая эти вопросы, заключается в том, что многие исследователи AI были прямо-таки шокированы моей идеей, что действительные феномены человеческого сознания могут зависеть от действительных физико-химических свойств действительных человеческих мозгов. Но если вы немного подумаете об этом, то поймете, что мне не следовало удивляться; ибо замысел сильного AI имеет хоть какие-то шансы на успех только, если вы принимаете некоторую форму дуализма. Замысел состоит в том, чтобы воспроизвести и объяснить ментальное, конструируя программы, но вы можете осуществить этот замысел лишь в том случае, если сознание не только концептуально, но и эмпирически не зависит от мозга, ибо программа совершенно не зависит от той или иной реализации. Вы можете надеяться воспроизвести ментальное, конструируя и запуская программы, лишь в том случае, если вы полагаете, что сознание отделимо от мозга как концептуально, так и эмпирически (сильная форма дуализма), ибо программы должны быть независимы от мозга, а равно и от любых конкретных форм инстанцииции. Если ментальные операции состоят в вычислительных операциях над формальными символами, то, следовательно, они никаким интересным образом не связаны с мозгом; единственная связь заключалась бы в том, что мозгу случилось быть одним из неопределенно большого числа типов машин, способных инстанциировать данную программу. Эта форма дуализма не совпадает с традиционной картезианской разновидностью, утверждающей, что есть субстанции двух родов, но это все же картезианский дуализм в том смысле, что он настаивает на том, что то, что есть в сознании специфически ментального, не имеет внутренней связи с действительными свойствами мозга. Этот дуализм, лежащий в основе AI, маскируется тем, что литература по AI содержит многочисленные инвек-

тивы против “дуализма”; чего эти авторы, по-видимому, не осознают, так это того, что предпосылкой их позиции является некая сильная версия дуализма.

“Может ли машина мыслить?” Я-то считаю, что только машины и могут мыслить, и в самом деле только очень особые виды машин, а именно мозги и машины, обладающие теми

же каузальными способностями, что и мозги. И это самое главное основание, почему сильный AI так мало рассказал нам о мышлении, ибо ему нечего сказать нам о машинах. По своему собственному определению, он касается программ, а программы — не суть машины. Чем бы еще ни была интенциональность, она биологический феномен, и ее бытие столь же вероятно, сколь оно каузально зависимо от таких конкретных биохимических особенностей ее происхождения, как лактация, фотосинтез и любые другие биологические феномены. Никому не придет в голову, что мы можем производить молоко и сахар, запустив компьютерную модель формальных последовательностей лактации и фотосинтеза, но когда заходит речь о сознании, многие люди упорно хотят верить в такое чудо по причине своего глубокого и прочно укорененного дуализма: сознание, которое они имеют в виду, зависит от формальных процессов и не зависит от совершенно конкретных материальных причин — в отличие от молока и сахара.

В защиту данного дуализма выражается надежда, что мозг — это цифровой компьютер (кстати, первые компьютеры часто называли “электронными мозгами”). Однако это не поможет. Конечно, мозг — цифровой компьютер. Раз любой предмет — цифровой компьютер, отчего бы мозгу не быть им. Но все дело в том, что каузальная способность мозга продуцировать интенциональность не может заключаться в том, что мозг инстанцирует некую компьютерную программу, ибо возьмите любую программу, и найдется такой предмет, который инстанцирует эту программу, но все же не имеет никаких ментальных состояний. В чем бы ни заключалось продуцирование мозгом интенциональности, оно не может заключаться в инстанцировании некоей программы, ибо никакая программа сама по себе недостаточна для интенциональности[6].

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

Fodor J. A. (1980). Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology // Behavioral and Brain Sciences, № 3, pp. 63-110.

McCarthy J. (1979). Ascribing Mental Qualities to Machines // Philosophical Perspectives in Artificial Intelligence / Ringle M (ed.), pp. 161—95. Atlantic Highlands, NJ: Humanities Press.

[Multiple authors] (1978). “Cognition and Consciousness in Non-Human Species” // Behavioral and Brain Sciences 1(4): entire issue.

Newell A. (1979). Physical Symbol Systems. Lecture at the La Jolla Conference on Cognitive Science Later published in Cognitive Science, 1980, № 4, pp 135-8J.

- and Simon H. A. (1963). GPS - A Program that Simulates Human Thought // Computers and Thought / Feigenbaum E. A., Feldman J. A. (eds.) pp 279-96 New York. McGraw-Hill

Pylyshyn Z. W. (1980). Computation and Cognition. Issues in the Foundation of Cognitive Science // Behavioral and Brain Sciences, № 3, pp. 111—32

Schank R. C., Abelson R. P. (1977). Scripts, Plans, Goals, and Understanding. Hillsdale, NJ: Erlbaum.

Searle J. R. (1979a). Intentionality and the Use of Language // Meaning and Use / Margolis A. (ed.) . Dordrecht: Reidel.

- (1979b). What is an Intentional State? // *Mind*, № 88, pp. 74-92. Weizenbaum J. (1965). ELIZA — A Computer Program for the Study of

Natural Language Communication Between Man and Machine // *Commun, ACM*, № 9, pp. 36-45.

— (1976). *Computer Power and Human Reason*. San Francisco: W. Freeman.

Winograd T. (1973). A Procedural Model of Language Understanding // *Computer Models of Thought and Language* / Schank R. C., Colby K. M (eds.), pp. 152—86. San Francisco: W. H. Freeman.

[1] *Searle J. Minds, Brains, and Programs* // *The Philosophy of Artificial Intelligence* / Boden M (ed.) Oxford, 1990. Перевод выполнен А. Л. Блиновым. Впервые статья была опубликована в журнале: “*The Behavioral and Brain Sciences*”, 1980, № 3, pp. 417- 424. © Cambridge University Press. — *Прим. ред.*

[2] Я, разумеется, не утверждаю, что сам Шэнк подписался бы под этими утверждениями.

[3] Мысленный эксперимент (нем.) — *Прим. перев.*

[4] Кроме того, “понимание” влечет как обладание ментальными (интенциональными) состояниями, так и истинность (правильность, успех) этих состояний. Для целей нашего обсуждения нас интересует лишь обладание этими состояниями.

[5] Интенциональность — это, по определению, то свойство определенных ментальных состояний, в силу которого они направлены на объекты и положения дел в мире или в силу которого они суть об этих объектах и положениях дел. Таким образом, полагания, желания и намерения суть Интенциональные состояния; ненаправленные формы тревоги и депрессии не являются интенциональными состояниями. Подробнее см. в *Searle* (1976b).

[6] Я обязан довольно большому числу людей обсуждением этих вопросов и их терпеливыми попытками победить мое невежество в искусственном интеллекте. Я бы хотел особенно поблагодарить Нада Блока, Хьюберта Дрейфуса, Джона Хогелэнда, Роджера Шэнка, Роберта Вилен-^ски и Терри Винограда