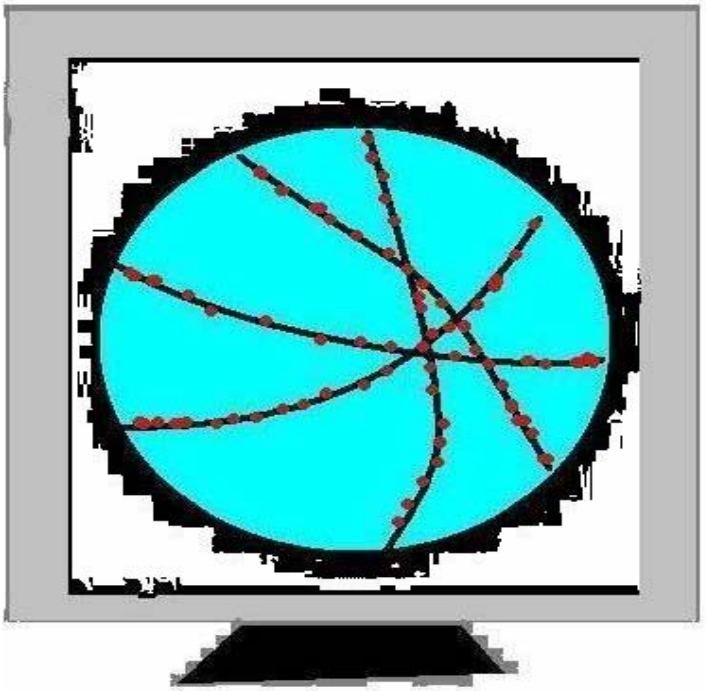

ИИнтелл

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ИНСТИТУТ ЭЛЕКТРОНИКИ И МАТЕМАТИКИ (ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ)

ТЕСТ ТЬЮРИНГА

РОБОТЫ ЗОМБИ

МОСКВА 2006

УДК 100.32
ББК 32.816
Т 96

Под редакцией к.ф.н. А.Ю. Алексеева

Т96 Тест Тьюринга. Роботы. Зомби. Перевод с англ. Под ред. А.Ю. Алексеева – М.: МИЭМ, 2006. – 120 с.

Книга знакомит с работами англо-американских философов – исследователей проблематики философии и методологии искусственного интеллекта. В ней отражается ряд вопросов, связанных с критическим осмыслением Теста Тьюринга (статьи Дж. Серля, Н. Блока, С. Брингсйорда, П. Беллоу, Д. Феруччи), с построением компьютерно-ориентированной онтологии «от первого лица» (работа А. Сломэна), с проблематикой философских зомби и принципом «несущественности сознания» (работа Т. Полджера, О. Фланагана).

Содержание книги может представлять интерес для философов, программистов, математиков, системологов, психологов, нейрофизиологов, лингвистов и всех тех, кого интересуют проблемы «дух/тело» («сознание/мозг»), теоретические основы когнитивных и компьютерных наук, методология искусственного интеллекта.

ISBN 5-94506-099-2

От редактора

Для подготовки рефератов и выступлений на семинарах по спецкурсу «Философия искусственного интеллекта» ко мне часто подходят студенты с просьбой помочь им в поиске литературы. К сожалению, приходится отвечать отрицательно – нет русскоязычной литературы, посвящённой данной проблеме. Почти нет литературы по методологии когнитивных и компьютерных наук. Крайне редко и малым тиражом издаются работы по философии сознания. А ведь эти темы – ведущие в мировой философии, точнее, в философии стран, вступающих в информационную эпоху.

Остаётся лишь одно советовать ребятам – пользуйтесь первоисточниками, переводите работы зарубежных авторов, раз нет крупных заделов в отечественной литературе. Следует отдать должное современному студенту – он достаточно бегло ориентируется в иностранных языках. Однако даже свободно владеющего иностранным языком студента подстерегают трудности, не преодолимые беглостью речи. У нас нет словарей по вышеуказанным темам, нет обзорных работ, учебных пособий, специальных монографий и пр.

Данный Сборник переводов, а также сборники, которые предполагается выпустить в ближайшем будущем, пускай незначительно, но позволят компенсировать острую потребность в материалах по философии искусственного интеллекта.

Следует отметить то, что работам в области философии искусственного интеллекта в 2005 году исполнился юбилей – 55 лет. Отсчёт принято вести с работы выдающегося английского логика и математика Алана Мейсона Тьюринга «Вычислительные машины и интеллект», опубликованной в журнале «Мышление и машины» в 1950 г. Небольшая по объёму статья сегодня получила громадное значение и стала наиболее цитируемой

работой в области ИИ. Представленные в ней идеи стали не только концептуальным каркасом ИИ, но и новейшей парадигмы философии сознания – функционализма, которая сегодня, в свою очередь, шагнула на порог социально-философской проблематики.

В настоящее время философия ИИ дисциплинарно не оформлена. Интересный вариант предлагает Давид Чалмерс. Философия ИИ рассматривается как подраздел (и весьма существенный) более общей философии – философии сознания.

Философия ИИ (Philosophy of Artificial Intelligence) по классификации Д. Чалмерса, включает следующие рубрики:

1.1. Могут ли машины мыслить?: 1.1a. Тест Тьюринга; 1.1b. Аргументы Гёделя (Лукас, Пенроус); 1.1c. Китайская комната (ключевая работа данной рубрики – статья Серля Джона Р. «Разумы, мозги, программы» (Перевод Д. Родионова) – опубликована в данном Сборнике); 1.1d. Сознание машины; 1.1e. Мышление машины.

1.2. Компьютация и репрезентация: 1.2a. Символы и символические системы; 1.2b. Вычислительная семантика; 1.2c. Имплицитные / эксплицитные правила и репрезентационизм; 1.2d. Арепрезентативный искусственный интеллект.

1.3. Философия коннекционизма: 1.3a. Коннекционизм и композициональность (Фодор/Пулушин); 1.3b. Репрезентация в коннекционизме.

1.3c. Коннекционизм и элиминативизм; 1.3d. Дискуссии между коннекционистами и «классиками» (репрезентативистами); 1.3e. Субсимволическая компьютация (Смоленский); 1.3g. Фундаментальные приложения философии ИИ.

1.4. Динамические системы.

1.5. Основные проблемы ИИ: 1.5a. Природа ИИ; 1.5b. Уровни анализа (Марр); 1.5c. Проблема фреймов; 1.5d. Методология ИИ.

1.6. Компьютационализм в когнитивных науках.

1.7. Компьютация и физические системы.

Общий объем книг и статей по философии ИИ (по вышеприведённой классификации Д. Чалмерса) составляет более 600 единиц.

Однако хочется с ним поспорить по поводу самой структуры классификатора. Сложно провести чёткую границу между философией ИИ и, к примеру, философией сознания. Авторы «пограничных» работ, раскрывая проблемы философии сознания, постоянно апеллируют к философии ИИ и, наоборот. Данный факт демонстрируют две работы, вошедшие в наш Сборник – Сломэна Арона «Что значит быть камнем?» (Перевод В. Крючкова), а также Фланагана Оуэна и Полджера Томаса «Зомби и функция сознания» (Перевод Т. Кураевой). Следует отметить, что последняя статья – статья по зомби – по сути является первой, достаточно крупной работой, ставшей доступной русскоязычному читателю.

Так же, к примеру, у Д. Чалмерса не нашла отражение проблема компьютерного творчества – в Сборнике этой проблеме посвящён перевод статьи Брингсйорда Селмера, Беллоу Пола и Феруччи Дэвида «Творчество, тест Тьюринга и улучшенный тест Лайвлейс» (Перевод А. Ласточкина). Неизвестно, к какому классу отнести работу Неда Блока «Психологизм и бихевиоризм» (Перевод И. Матанцевой и И. Чижова) – в ней просматривается многообразие тестов Тьюринга в контексте частных парадигм философии сознания.

Таким образом, видится иная форма классификации, более гибкая, неиерархическая. Предпочтительнее, конечно, фасетная классификация. Однако выявление оснований новой классификации упирается в необходимость перевода работ по философии ИИ.

Хочется выразить уверенность, что начало систематическим переводам в области философии ИИ положено в настоящем Сборнике.

Андрей Алексеев

Джон Р. Серль

РАЗУМЫ, МОЗГИ, ПРОГРАММЫ

Перевод Дениса Родионова

Перевод выполнен по работе: Searle, John R., «Minds, Brains, and Programs», The Behavioral and Brain Sciences, vol. 3, 1980 (http://www.cs.ucsb.edu/~cs165a/articles/Searle.html)

Как с философской и психологической позиций следует трактовать попытки компьютерного моделирования когнитивных способностей? При ответе на этот вопрос целесообразно различать *сильный* искусственный интеллект (ИИ) и *слабый* (или «осторожный») ИИ. Что касается слабого ИИ, то роль компьютера в изучении разума очевидна: компьютер выступает здесь весьма мощным инструментом. Например, становится возможным более строго и точно формулировать и проверять гипотезы. Но в концепции сильного ИИ компьютер – отнюдь не средство изучения разума. В этом случае должным образом запрограммированный компьютер, собственно, и есть разум – т.е. посредством соответствующей программы компьютер в буквальном смысле слова *понимает* и обладает иными когнитивными способностями. А поскольку компьютеру становятся присущи когнитивные состояния, то оказывается бессмысленным использовать его программы для подтверждения психологических объяснений – программы сами становятся подобными объяснениями.

У меня нет возражений по поводу теории о слабом ИИ, по крайней мере, в рамках данной статьи. Моё внимание будет направлено на положения сильного ИИ, особенно на утверждения, что соответствующим образом запрограммированный компьютер в прямом смысле реализует когнитивные функции и что таким образом программы объясняют способность человеческого понимания. Ниже при упоминании ИИ я буду подразумевать именно сильный его вариант, соответствующий этим двум утверждениям.

В основе моей работы лежат исследования Роджера Шэнка и его коллег из Йельского университета (Шэнк и Абельсон, 1977). С этими работами я знаком лучше, чем с другими изысканиями в данной области. Они весьма наглядно раскрывают предмет моего изучения. Моя аргументация может применяться не только к программе Шэнка, но и к таким программам, как SHRDLU Винограда (Виноград, 1973), ELIZA Вейзенбаума и, в принципе, к любой машине Тьюринга, моделирующей психические феномены.

Кратко, не вдаваясь в подробности, программу Шэнка можно охарактеризовать следующим образом. Она предназначена для моделирования человеческой способность понимать тексты. За достоверный факт понимания текста принимается способность отвечать на вопросы о его содержании, в том числе в условиях неявной представленности в тексте запрашиваемой информации. Предположим, перед вами текст: «Человек зашёл в ресторан и заказал гамбургер. Когда гамбургер принесли, он оказался зажаренным до угольков, и человек в ярости вылетел из ресторана, не заплатив за гамбургер и не дав официанту на чай». Вас спрашивают: «Съел ли человек гамбургер?» Вы, скорее всего, ответите: «Нет, не съел». Подобным образом вам предлагается текст: «Человек зашёл в ресторан и заказал гамбургер. Когда гамбургер принесли, человек остался им доволен. Покидая ресторан, он прежде, чем заплатить по счёту, дал официанту на чай». Вас снова спрашивают: «Съел ли человек гамбургер?» И вероятно вы ответите: «Да, конечно, съел». Машины Шэнка способны отвечать на вопросы подобного рода. Для реализации данной способности они обладают «репрезентацией» информации такого же типа, как и у людей, рассуждающих о ресторанах и тому подобном. Когда в машину вводят текст и затем задают вопрос, она выдаёт ответы, которые при тех же условиях диалога ожидается получать и от людей. Сторонники сильного ИИ утверждают, что в этой последовательности вопросов и ответов машина не просто моделирует человеческую способность, но также, во-первых, способна в прямом смысле *понимать* тексты и отвечать на вопросы, и, во-вторых, машина и её программа *объясняют* человеческую способность понимать текст и отвечать на вопросы о нем.

Оба эти утверждения, однако, никак не обосновываются работой Шэнка. Я попытаюсь показать это ниже. При этом я не буду отрицать, что [машины Шэнка] в принципе имеют отношение к данным утверждениям.

Один из способов проверить любую теорию разума – это спросить себя, что было бы, если мой собственный разум работал бы так же, как, согласно теории, работают все разумы. Проверим программу Шэнка в следующем *мысленном эксперименте*. Предположим, я заперт в некоторой комнате. Передо мной – набор листов с китайскими письменами. Предположим теперь (и это будет соответствовать действительности), что китайского языка – ни письменного, ни разговорного – я не знаю. Я даже не могу отличить китайские иероглифы, скажем, от японских или любых других бессмысленных сочетаний завитков. Китайские письмена как раз и являются для меня не более чем бессмысленными завитками. Далее предположим, что в добавление к первому набору листов с иероглифами мне дали второй такой же и к нему – приложение в виде списка правил соответствия наборов друг другу. Правила составлены на английском языке, и я понимаю их так же хорошо, как может их понять любой носитель английского языка. Они позволяют мне соотносить одну совокупность формальных

символов с другой («формальность» означает здесь способность различения иероглифов по форме). Предположим затем, что мне предоставили ещё один, третий, набор листов с китайскими иероглифами. К нему также прилагается инструкция на английском языке, позволяющая соотносить содержимое третьего набора с содержимым двух предыдущих, т.е. сопоставлять определённые формальные типы китайских иероглифов первых двух наборов с другими типами иероглифов, записанных на листах третьего набора. И мне совершенно неведомо, что люди, предоставившие мне листы с иероглифами, называют первый набор «алфавитом», второй – «текстами», третий – «вопросами». Также мне неизвестно и то, что иероглифы, выдаваемые мной в форме реакции на иероглифы третьего набора, называют «ответами на вопросы», а список правил на английском – «программой». Теперь усложним эксперимент, допустив, что параллельно люди дают мне тексты на английском, которые я отлично понимаю. Меня по-английски спрашивают о содержании этих английских текстов, и я отвечаю им также на английском языке.

Спустя некоторое время я настолько привыкаю следовать инструкциям по манипулированию китайскими иероглифами, а программисты так быстро пишут соответствующие программы, что с точки зрения стороннего наблюдателя, находящегося вне комнаты, в которой я заперт, мои ответы на вопросы становятся совершенно неотличимы от ответов китайца. И глядя на мои ответы, никто не сможет предположить, что я ни слова не понимаю по-китайски. Несомненно, что мои ответы на английские вопросы неотличимы от ответов, которые мог бы выдать любой англичанин, по той простой причине, что я сам являюсь носителем английского языка. И для стороннего наблюдателя, читающего мои «ответы», одинаково хорошими представляются как ответы на английские вопросы, так и на китайские. Но в отличие от вопросов на английском языке, в китайском варианте диалога я ничего не понимаю, а просто реагирую посредством манипулирования формальными символами, интерпретируемыми [сторонним наблюдателем]. В отношении текстов на китайском языке я действую подобно компьютеру – осуществляю вычислительные операции над формально определёнными элементами. В этом случае я сам являюсь реализацией компьютерной программы.

Вспомним утверждения сильного ИИ о том, что соответствующим образом запрограммированный компьютер понимает тексты и что программы в некотором смысле объясняют человеческую способность понимания. Изучим их в свете нашего мысленного эксперимента.

1. Рассмотрим первое утверждение сильного ИИ. Очевидно, что в нашем эксперименте я ни слова не понимаю по-китайски. Моя входная и выходная информация может быть совершенно неотличима от той информации, которая имеется у носителя китайского языка. Также у меня может

быть программа [, реализующая инструкции,] на любой самый взыскательный вкус. Однако я по-прежнему ничего не буду понимать. По тем же причинам и компьютер Шэнка ничего не понимает в текстах – китайских, английских и любых других. В случае с китайскими текстами компьютер – это я, ничего не понимающий по-китайски. В случае с английскими текстами я уже не выступаю в роли компьютера, а у него нет ничего сверх того, что имеется у меня, когда я ничего не понимаю.

2. Рассмотрим второе утверждение сильного ИИ – о способности программы объяснять человеческую способность понимания. Мы видим, что компьютер и его программа не дают достаточных условий для понимания. Компьютер и программа просто функционируют, а понимания – нет. Может быть, они задают необходимые условия понимания, внося тем самым существенный вклад в объяснение данной когнитивной способности? Одно из утверждений сторонников сильного ИИ состоит в том, что при понимании текстов по-английски, я проделываю в точности то же самое, – может быть, чуть больше того, – что и при манипулировании китайскими иероглифами. Просто в случае с английским, когда понимание налицо, в моём распоряжении больше формальных символов, чем в случае с китайским, когда понимание отсутствует. Я не стал демонстрировать ложность этого суждения. Данный аргумент предполагает, что мы можем написать программу, у которой входы и выходы будут такими же, как и у носителя языка. Но, кроме того, допускается, что носители языка – в силу ряда характерных признаков – суть экземпляры программы. На основе этих двух аргументов можно допустить, что даже если программа Шэнка и не рассказывает нам историю понимания до конца, то, по крайней мере, делится фрагментами этой истории. Возможно, я и могу допустить эмпирическую возможность понимания [компьютером текстов]. Однако в свете нашего примера оно совершенно не просматривается. Я продемонстрировал (может и недостаточно ясно), что моё человеческое понимание текстов никак не соотносится с «пониманием» текстов компьютерной программой. В случае китайского у меня есть всё, что искусственный интеллект может предоставить посредством программы. При этом я совершенно ничего не понимаю. В случае же с английским я всё понимаю, но при этом нет повода предполагать, что моё понимание каким-то образом связано с компьютерными программами, выполняющими вычислительные операции над формально определёнными элементами. Пока программа определяется в терминах вычислительных операций над элементами, задающимися чисто формальным способом, наш пример показывает, что сами по себе программы никак не связаны с пониманием. Мы видим, что это, безусловно, недостаточные условия и у нас нет никаких оснований считать их достаточными или дающими хоть сколько-нибудь достойное объяснение способности понимать. Заметим, что сила аргумента заключается не в том, что имеется возможность достичь функциональной тождественности в способах понимания посредством применения различных формальных правил (то есть когда одинаковы входы и выходы различных программ). Напротив, суть в другом: какие бы формальные правила вы не вложили в компьютер, их будет недостаточно для объяснения способности понимать, ведь человек может следовать формальным правилам без всякого их понимания. Нет никаких оснований для утверждения о том, что подобные правила необходимы или хотя бы полезны, ведь нет причин предполагать, что для понимания английского я должен оперировать какими-то формальными программами.

Но тогда что же это такое, что у меня имеется в случае с английскими текстами и чего у меня нет в случае с китайскими? Ответ очевиден: это смысл. В первом случае мне понятен смысл текстов. Во втором случае он абсолютно мне неизвестен. Но что такое смысл и почему для машины он непостижим? К данному вопросу я вернусь позже, сейчас же продолжу анализ эксперимента.

О своём мысленном эксперименте я рассказал исследователям искусственного интеллекта и получил от них много отзывов. Любопытно то, что мнения разошлись. Ниже я рассмотрю типовые отзывы, указав также географическое место работы исследователей.

Вначале следует разъяснить заблуждения относительно слова «понимание». О его значении можно спорить сколь угодно долго. Мои критики указывают, что «понимание» - это многоуровневая структура, а не просто двуместный предикат; что имеются различные способы и формы понимания; что [логический] закон исключенного третьего напрямую неприменим к утверждениям типа «*x* понимает *y*»; что во многих случаях от субъективного решения, а не от объективных факторов зависит, что *x* понимает *y*, и т.д. На всё это я спешу ответить: конечно, вы правы… Но к нашему спору это не относится. Можно достаточно чётко очертить ситуации, в которых слово «понимание» употреблять уместно, и ситуации, в которых употреблять его совсем не нужно. Различение этих ситуаций важно для цели нашего обсуждения[1]. Я понимаю тексты на английском языке, в меньшей степени – на французском, в ещё меньшей степени – на немецком. Тексты на китайском я не понимаю вовсе. С другой стороны, мой автомобиль и мой арифмометр не понимают ничего – и это тот случай, когда термин «понимание» применять не следует. Однако «понимание» и другие когнитивные предикаты посредством метафор и аналогий часто применяются к артефактам – автомобилям, арифмометрам и пр. Но при таком употреблении термина он ничего не обозначает. Например, привычно звучат выражения: «Дверь *знает*, когда открыться, у неё есть фотоэлементы»,

[1] «Понимание» подразумевает как наличие психических (интенциональных) состояний, так и оценку истинности (правильности, эффективности) данных состояний. Мы будем рассматривать только наличие состояний (*прим. авт.*)

«Счётная машина *знает* (*понимает, умеет*), *как* складывать и вычитать, но *не знает*, как делить», «термостат *чувствует* изменения температуры». Причина подобных присваиваний весьма любопытна. Она обусловлена перенесением на артефакты нашей собственной интенциональности.[2] Так как наши инструменты – это продолжения наших замыслов, то нам кажется вполне естественным метафорически присваивать им свойства интенциональности. Однако подобные примеры «с пониманием» не расколют философский лёд непонимания [вольности употребления слов]. «Смысл понимания» автоматической дверью «инструкций» фотоэлемента – это совсем не тот смысл, в котором я понимаю английские тексты. Рассуждения о «смысле» корректны, если «смысл» понимания текстов программируемыми компьютерами Шэнка отождествляется со «смыслом», в котором «понимает» дверь, и противополагается «смыслу» понимания английского языка. Однако у некоторых исследователей ИИ иное мнение. Так, Ньюэл и Саймон (1963) утверждают о тождественности способа понимания, предложенного ими для компьютерной реализации, человеческому способу понимания. Мне нравится прямолинейность их суждений. Утверждения именного такого типа я и буду изучать. Я утверждаю, что запрограммированный компьютер в прямом смысле понимает то же, что понимают автомобиль и счётная машина. А именно: не понимает ничего. Понимание компьютера не является (подобно моему пониманию немецкого) частичным или неполным. Оно отсутствует.

А теперь приступим к анализу отзывов на наш эксперимент.

1. Системологический отзыв (Беркли)

«Мы согласны, что индивид, запертый в комнате, действительно не понимает тексты. Однако следует учесть, что он лишь часть системы. Система же в целом тексты понимает. Индивид в рамках системы не изолирован. Перед ним толстая книга со сводом всевозможных правил. У него много черновиков и ручек для осуществления вычислений. Он хранит и обрабатывает китайские йероглифы на основе использования банков данных. В таких условиях понимание невозможно приписать лишь отдельному индивиду, его следует приписать всей системе, частью которой индивид является».

Мой ответ на системологический подход прост. Предоставьте индивиду включить в себя все элементы системы. Он запомнит все инструкции из книги правил и все китайские йероглифы из базы данных. Вычисления он будет осуществлять в уме. Такой индивид будет олицетворять целостную систему и у последней невозможно будет выявить такие функции, которые индивид бы не выполнял, и такие ресурсы, которыми бы он не обладал. Мы даже можем избавиться от комнаты – пусть индивид работает на свежем воздухе. Но в любом случае он по-прежнему ничего не понимает по-китайски, и тем более не понимает по-китайски система, ведь в ней нет ничего сверх того, что есть в индивиде. Если не понимает индивид, то и система также не понимает по той причине, что она – лишь часть индивида.

2 Интенциональность – это свойство определённых психических состояний, в соответствии с которым они обязательным образом направлены либо на конкретные объекты, либо на мир в целом. Убеждения, желания и намерения являются интенциональными состояниями. Ненаправленные формы, например, волнения и депрессии, таковыми не являются (*прим. авт.*)

На самом деле я не берусь всерьёз оппонировать системологическому отзыву, так как системологическая теория [в контексте нашего рассуждения] представляется мне изначально порочной. Я имею в виду идею о том, что хотя индивид и не понимает китайского, но некое *соединение* индивида с листами бумаги китайский понять может. Мне сложно вообразить, что кто-то также может посчитать эту идею правдоподобной, если, конечно, он не поклонник системологической методологии. Однако, по-моему, сторонники сильного ИИ в конечном итоге склоняются к подобным системологическим мыслям. Поэтому я более развёрнуто отвечу на системологический отзыв. В одной из его версий приводилось следующее: хотя человек, олицетворяющий систему в целом, не понимает китайского языка в том смысле, в котором китаец понимает по-китайски (потому что, к примеру, он не знает, что в тексте говорится о ресторанах, гамбургерах и т.д.), всё же «человек как система манипулирования формальными символами» *реально понимает китайский*. При этом подсистема человека, обрабатывающая формальные символы китайского языка, не пересекается с подсистемой, которая обрабатывает символы английского языка.

То есть, по сути, человек как система включает в себя две подсистемы. Одна подсистема понимает английский, вторая – китайский. Данные «подсистемы слабо взаимодействуют друг с другом» [,как мог бы утверждать сторонник системного подхода]. Хочется добавить, что они не только слабо взаимодействуют, они принципиально различны. Подсистема, понимающая английский (допустим на некоторое время системологическую терминологию), знает, что тексты сообщают о ресторанах и поедании гамбургеров. Она знает, что её спрашивают о ресторанах и что ответы на вопросы станут весомее, если она будет заимствовать данные из контекста, и т.д. Китайская подсистема ничего этого не знает. Там, где английская подсистема знает, что референтом слова «гамбургеры» выступают реальные гамбургеры, китайская подсистема знает лишь то, что за «штрих-штрих» следует «чёрк-чёрк». Всё, что знает китайская подсистема, это формальные символы, поступающие на вход подсистемы, правила их обработки, записанные на английском языке, и формальные символы на выходе подсистемы. Цель нашего эксперимента и состояла в показе того, что само по себе манипулирование символами ни в каком смысле не может быть достаточным для понимания китайского – ведь человек может написать «чёрк-чёрк» после «штрих-штрих» без всякого понимания китайского

языка. В силу таких соображений очевидна и неприменимость аргумента выделения в человеке подсистем: подсистемы не могут понимать больше человека, их объединяющего. В подсистемах по-прежнему нет ничего, что хотя бы отдалённо напоминало способ понимания англо-говорящим человеком (английской подсистемой) текстов. Китайская подсистема – это просто часть английской подсистемы, бессмысленно обрабатывающая символы согласно правилам на английском языке.

Зададимся вопросом: что в нашей ситуации оправдывает системологический отзыв, т.е. какие *объективные* основания поддерживают суждения о необходимости включения в состав индивида подсистемы, понимающей (в прямом смысле слова «понимание») китайские тексты? По-моему, единственное объективное основание, применимое и к моему примеру, это та же самая входная и выходная информация, которая, как предполагается, имеется и у носителя китайского языка, а также некая программа переработки входной информации в выходную. Но цель примера состояла в стремлении показать, что этого отнюдь не достаточно для возможности понимания в том смысле слова, в каком я пониманию тексты на английском языке. Индивид и, следовательно, все подсистемы, совместно его образующие, могут обладать требуемой [с точки зрения китайца] совокупностью: входом, выходом, программой. Но все эти подсистемы и индивид в целом по-прежнему не будут способны понимать так же, как я понимаю английский. Единственным поводом говорить о *необходимости* наличия во мне подсистемы понимания китайских текстов является программа, способствующая прохождения мною теста Тьюринга. И это позволяет мне обмануть носителей китайского языка. Строго говоря, одним из базовых положений нашей работы выступает адекватность [моего поведения] тесту Тьюринга. Пример показывает, что возможны две «подсистемы». Обе пройдут тест Тьюринга, но только одна из них будет понимать. На первый взгляд, истинно суждение: обе подсистемы проходят тест Тьюринга, поэтому они обе понимают. Однако очевиден наш контраргумент: моя подсистема, понимающая английский язык, заключает в себе много больше, нежели чем та подсистема, которая просто обрабатывает китайские символы. После атаки нашего аргумента системологический отзыв начинает молить о пощаде, так как он совершенно неаргументированно утверждал, что если не часть, то система в целом понимает по-китайски.

При рассмотрении системологического отзыва с другой стороны можно увидеть, что он приводит к иным абсурдным выводам. Если принимается утверждение, что мне присуще понимание на основании наличия у меня входной и выходной информации определённого типа, а также программы между входом и выходом, то все некогнитивные подсистемы становятся когнитивными. Например, можно составить такое описание функционирования желудка, характеризующее обработку желудком информации, которое позволяет говорить о том, что желудок «понимает» (ср. Пу-

люшин, 1980). Подобного рода концепции иллюстрирует множество компьютерных программ. Однако вряд ли можно согласиться с такими суждениями. И если мы соглашаемся с системологическим отзывом, то становится сложно избежать утверждений, что желудок, сердце, печень и т.п. – суть понимающие подсистемы, так как нет критерия отличения мотива суждений о «понимании» китайских подсистем от мотива суждений по поводу желудочного «понимания». Возникает неразбериха, подобная следующему. У китайской подсистемы имеется входная и выходная информация, и у желудка имеется вход (пища) и выход (продукты желудочной обработки). С точки зрения действующего субъекта – с моей точки зрения – информацию невозможно обнаружить ни в пище, ни в китайском языке. Китайский язык – это просто совокупность ничего не значащих завитков. Информация не содержится ни в них, ни в пище. Она существует исключительно в головах программистов и интерпретаторов. Однако при определённой силе воображения ничто не может удержать программистов и интерпретаторов от произвольного представления веществ на входе и выходе пищеварительной системы в виде информации.

Приведённое соображение связано с конкретными проблемами сильного ИИ. Они крайне важны, хотя напрямую и не связаны с нашим экспериментом. Для их изучения немного отвлечёмся. Если сильный ИИ претендует на роль психологической концепции, он должен выработать критерий отличения ментальных систем от нементальных. Для этого он должен быть способен различать принципы, в соответствии с которыми разумное функционирование системы можно отличить от функционирования неразумного. В противном случае сильный ИИ не даст нам ответа, в чём же состоит специфика ментальности. При этом дистинкция «ментальное/нементальное» не может базироваться только на внешних наблюдениях. Она должно стать и внутренним определением системы, иначе наблюдатель начнёт произвольно выносить суждения, например, о неспособности людей к мышлению, или, скажем, о том, что ураганы мыслят. В работах по ИИ данная дистинкция настолько расплывается, что становится сомнительным отнесение искусственного интеллекта к сфере интеллектуальных исследований. Маккарти, к примеру, пишет: «Можно утверждать, что машины (для простоты примера, термостаты) обладают убеждениями. Наличие у машин убеждений – это характеристика большинства машин, предназначенных решать проблемы» (Маккарти, 1979). Любому, кто полагает, что сильный ИИ – это теория разума, следует считаться со следствиями, вытекающими из данного замечания. Нас просят согласиться (как с [научным] открытием в области сильного ИИ) с тем, что термостат, этот кусок металла, предназначенный для регулирования температуры, обладает убеждениями той же природы, которыми обладают наши супруги и дети. Более того, сильный ИИ призывает согласиться, что убеждения (в прямом смысле слова «убеждение») имеются у большинства других при-

боров в моей комнате: у телефонов, магнитофонов, арифмометров, выключателей. Не буду оспаривать точку зрения Маккарти (это не является целью моей работы) и нижеследующее высказывание приведу без доказательств. Изучение разума начинается с признания факта наличия у человека убеждений, в то время как у термостатов, телефонов и арифмометров убеждений нет. Если у вас имеется опровержение, то приведите хоть какой-нибудь аргумент! Создаётся впечатление, что исследователи ИИ, пишущие подобные вещи, позволяют себе вольное отношение к предмету изучения в силу несерьёзного к нему отношения и полагают, что и другие столь же легкомысленны. Предлагаю хотя бы ненадолго воспринимать предмет должным образом. Напрягите мозг и подумайте о том, как приписать «убеждения» куску металла на стене. Я имею в виду убеждения с пропозициональным содержанием и состояниями удовлетворения, изменяющиеся в зависимости от условий; убеждения страстные и слабые; нервозные, тревожные и безмятежные; догматические, рациональные и суеверные; слепую веру и убеждения, постоянно колеблющиеся; какие угодно иные. С термостатом ничего не выйдет – убеждения ему приписать невозможно. Также не выйдет и с желудком, печенью, арифмометром и телефоном. И поскольку мы договорились быть серьёзными, заметим, что очевидная истинность последнего суждения опровергает претензии сильного ИИ быть наукой о разуме. Разум не пребывает повсюду. Мы стремились понять, чем разум отличается от термостатов и печени. И даже если бы Маккарти был прав, то сильный ИИ не мог бы нам об этом поведать.

2. Робототехнический отзыв (Йель)

«Предположим, разработана программа, отличающаяся от программы Шэнка, и компьютер с новой программой – это «внутренности» робота. Отличается она тем, что не только получает формальные символы на входе и выдаёт формальные символы на выходе, но, помимо этого, управляет роботом. Благодаря программе робот воспринимает окружающее, передвигается, гуляет, забивает гвозди, ест, пьёт – всё, что угодно. Чтобы «видеть», на корпусе робота установлена телевизионная камера. У него есть «руки» и «ноги». Ими робот производит «действия». Всё это контролируется компьютерным «мозгом». Такой робот, в отличие от компьютера Шэнка, обладает подлинным пониманием и другими ментальными свойствами».

Первое, что следует отметить в робототехническом отзыве – это молчаливое признание того, что когнитивная способность не сводится исключительно к способности манипулирования формальными символами. Помимо этого, когнитивная способность обусловлена совокупностью самых различных взаимосвязей с окружающим миром (ср. Фодор, 1980). Однако добавление «перцептивных» и «двигательных» способностей не добавляет ничего нового ни по поводу понимания в частности, ни по поводу

интенциональности в целом. В этом смысле новая программа ничем не отличается от программы Шэнка. Для доказательства проведём наш мысленный эксперимент для случая с роботом. Допустим, внутри робота вместо компьютера нахожусь я. В отличие от исходного примера, мне предоставляется большее количество китайских иероглифов и инструкций на английском по сопоставлению одних иероглифов с другими и выдаче китайских иероглифов. Без моего ведома часть китайских иероглифов поступает ко мне от телевизионной камеры, присоединённой к роботу, а другая часть иероглифов, выдаваемых мною, управляет двигателями робота, перемещая его руки и ноги. Важно подчеркнуть, что все мои действия сводятся к манипулированию формальными символами. Иные обстоятельства мне неизвестны. Я получаю «информацию» от перцептивного аппарата робота и даю «инструкции» его двигательному аппарату, не зная об этих аппаратах. Я гомункул робота, но в отличие от традиционных гомункулов, я не ведаю, что происходит. Я ни о чем не знаю, помимо правил манипуляции символами. Следует добавить, что роботу не присущи никакие интенциональные состояния, он просто перемещается в соответствии с программой и электрическими импульсами в проводах. В свою очередь, руководствуясь программой, я так же не обладают никакими интенциональными переживаниями. Всё, что я делаю, это попросту следую формальным инструкциям по манипулированию формальными символами.

3. Отзыв с позиции моделирования мозга (Беркли и Массачусетский технологический институт)

«Предположим, разработана программа, которая основана не на способе репрезентации информации о мире, как скрипты Шэнка, а на моделировании синапсических связей при возбуждении нейронов мозга китайца в моменты восприятия и понимания им текстов на китайском языке, а также при ответах на связанные с тестами вопросы. Машина воспринимает тексты и вопросы о них на китайском языке, моделирует строение китайского мозга (т.е. мозга, принадлежащего носителю китайского языка) при обработке таких текстов и выдаёт ответы на китайском языке. Можно также представить, что при работе с родным языком машина управляется не одной последовательной программой, а набором параллельных программ, что в большей мере соответствует гипотезам о принципах функционирования настоящего мозга. Тогда с уверенностью можно заявить: машина понимает тексты. Если это отрицать, то следует отрицать и возможность понимания китайцами китайских текстов. Разве можно отличить программу, протекающую на уровне синапсов в мозгу китайца при понимании им китайских текстов, от компьютерной программы, осуществляющей подобного рода понимание»?

Прежде чем возразить на данный отзыв, следует заметить, что он достаточно странен для защитника искусственного интеллекта (или сто-

ронника парадигмы функционализма, что, по сути, то же самое). Ранее я думал, что сущность сильного ИИ как раз заключается в том, что для знаний о работе разума не требуется знаний о работе мозга. Как мне казалось до этого отзыва, основная гипотеза ИИ заключается в существовании особого уровня ментальных операций, на котором конституируется сущность разумного. Эти ментальные операции – суть вычислительные процедуры над формальными элементами. Данный уровень может быть реализован при участии мозговых процессов самого различного типа, подобно тому, как компьютерная программа может быть реализована посредством разных аппаратных средств. По основной формуле сильного ИИ разум относится к мозгу так же, как программа относится к аппаратным средствам – можно изучать работу разума, не вдаваясь в нейрофизиологические исследования. Однако рассмотрение деятельности мозга в таком приближении недостаточно для уяснения того, как возможно понимание. Убедимся в этом. Представим, что человек, сидящий в нашей «китайской» комнате и ничего не понимающий по-китайски, перебирает не символы, а управляет системой водопроводных труб, соединённых клапанами. Когда человек получает китайский иероглиф, он ищет в программе, написанной на английском, какие клапаны ему нужно открыть, а какие – закрыть. Каждое соединение водяных отсеков соответствует синапсу в мозгу китайца, а система в целом устроена таким образом, что после должной последовательности возбуждений (что в свою очередь соответствует должной последовательности изменений положений вентилей) ответ на китайском языке вытекает из труб, соединённых требуемым образом.

Но где же в этой системе понимание? На входе система труб принимает китайские тексты и вопросы, и, моделируя формальную структуру мозга китайца, на выходе выдаёт ответы на китайском языке. Но человек определённо не понимает китайских текстов. Не понимают их и водопроводные трубы. Если же посчитать, что понимать способна *конъюнкция* человека *и* водопроводных труб (что я считаю совершенно абсурдным), то следует заметить, что, в принципе, человек может представить формализованную структуру водопроводных труб и в своём воображении произвести все «нейронные возбуждения». Проблема с моделированием мозга в том, что моделируются не то, что нужно. Моделируется лишь формализованная последовательность возбуждений нейронов посредством синапсических связей, а важно смоделировать каузальные, т.е. причинные зависимости [, определяющие соотношение «мозг-сознание»]. Именно причинные зависимости задают возможность интенциональных переживаний. Формальных зависимостей для этого недостаточно. Данный факт и демонстрирует пример с водопроводными трубами: мы можем полностью отделить формальные свойства от важных для возможности понимания нейробиологических причинных зависимостей.

4. Комплексный отзыв (Беркли и Стэнфорд)

«Каждый из трёх предыдущих отзывов в отдельности не убедителен для опровержения аргумента «китайской комнаты». Более убедительным и даже решающим оказывается объединение отзывов, когда все они приводятся в комплексе. Представим робота, компьютер которого посредством соответствующей программы формально воспроизводит все синапсические связи человеческого мозга. Пусть поведение робота будет неотличимо от поведения человека. Представим всё это в целом, как единую систему, а не просто в виде отдельного компьютера со своим входом и выходом. Вне всяких сомнений, данной системе можно приписать интенциональность».

Я целиком и полностью согласен, что гипотеза об интенциональности робота выглядит убедительно, с ней трудно не согласиться. Однако в комплексном отзыве существенными является только внешние проявления поведения. Все остальные компоненты неуместны. Так представление компьютерного мозга как формального аналога человеческого мозга излишне. Если бы удалось построить робота, чьё поведение на протяжении достаточно длительного времени не отличалось бы от поведения человека, то такому роботу можно было бы приписать интенциональность. Однако всё это – до поры до времени, пока не обнаружились бы отличия в поведении робота, которые позволили бы судить о неуместности такого приписывания роботу интенциональности.

На самом деле я не вижу, как комплексный отзыв может поддержать утверждения сильного ИИ. И вот почему. В соответствии с сильным ИИ, реализация формальной программы с соответствующими входом и выходом – это необходимое и достаточное условие интенциональности. Как утверждает Ньюэлл (1979), сущность ментального заключена в способах оперирования системой физических знаков. Но интенциональность, которую мы сейчас пытаемся приписать роботу, не имеет ничего общего с его способностью оперировать формальными программами. Это приписывание основано на простом утверждении, что если робот выглядит, как человек, и его поведение в достаточной степени похоже на человеческое, то можно предположить, что, пока не доказано обратное, робот способен пребывать в ментальных состояниях, подобных человеческим. Эти состояния выражаются в поведении робота и оказывают на него влияние. Также у робота должен быть внутренний механизм, способный вызывать такие ментальные состояния. Если бы мы знали, как объяснить поведение робота вне зависимости от подобных предположений, то мы не стали бы приписывать ему интенциональность, особенно если известно, что его поведение обусловлено реализацией формальной программы. И это в точности совпадает с вышеизложенным выше ответом на второе возражение.

Допустим, известно, что поведение робота целиком объясняется фактом получения человеком, сидящим у него внутри, формальных симво-

лов от сенсорных датчиков и отправки формальных символов в двигатели. Символы при этом человеком не интерпретируются, он манипулирует ими в соответствии с определённой совокупностью правил. Далее предположим, что человек ничего не знает о роботе, внутри которого он сидит. Все, что он знает, это какие операции с бессмысленными символами ему следует осуществлять. В таком случае робот – это искусная механическая кукла. Гипотеза же о том, что кукла обладает разумом, оказывается неподтверждённой и ненужной – ведь нет причин приписывать интенциональность роботу или системе, частью которой он является (исключая, конечно, человеческую интенциональность при манипулировании символами). Манипулирование формальными символами производится, входы и выходы корректны и соответствуют требованиям, однако основной носитель интенциональности – это человек, а ему неведомы интенциональные переживания (связанные с внешними проявлениями поведения). Например, человек не *видит*, что поступает через «глаза» робота, он не *стремится* двигать «рукой» робота, он не *понимает* никаких замечаний со стороны робота или роботу адресуемых. Также по ранее объяснённым причинам к интенциональности не способна и система, частями которой являются и человек и робот.

Чтобы в этом убедиться, противопоставим данную ситуацию случаям, в которых люди естественным образом приписывают интенциональность, например, приматам или домашним животным, соответственно, обезьянам или собакам. Причин для этого, грубо говоря, две: без подобного приписывания невозможно понять намерений животных и очевидно, что звери в чем-то похожи на людей (это у них – глаз, это – нос, а вот это – кожа и т.п.). Когерентность поведения животных человеческому поведению, а также вера в единство природы выступают обоснованиями для целого ряда допущений: наличия ментальных состояний у животных; обусловливания ими поведения; продуцирования ментальных состояний животных теми же механизмами, какими порождаются ментальные состояния человека. Подобные допущения применимы и к роботу. Однако мы их не признаём, поскольку знаем, что поведение робота – это результат работы формальной программы и что роботу не присуща та физическая субстанция, на основе которой порождаются реальные причинные зависимости. Поэтому роботу следует отказать в интенциональности.

Ниже рассматриваются ещё два типовых отзыва по поводу моего эксперимента. Они высказывались весьма часто и потому достойны рассмотрения. Правда, они немного уводят нас от основной темы.

5. Отзыв с позиции проблемы «другого сознания» (Йель)

«Откуда известно, что другие люди понимают китайский язык, и откуда известно, что другие люди вообще что-либо понимают? Только исходя из анализа поведения. Компьютер в принципе способен проходить

тесты [на когнитивные способности] с такими же результатами, как у людей. Поэтому, если приписывать способность понимания другим людям, то её следует приписывать и компьютерам».

Данное возражение заслуживает лишь краткого ответа. Проблема состоит не в том, каким образом мне становится известно, что у других имеются когнитивные функции, а в том, что же в действительности я им приписываю, говоря об этих функциях. Мой аргумент состоит в том, что когнитивные функции не могут быть представимы в форме вычислительных процессов и их внешних проявлений – вычислительные процессы и их проявления могут существовать без каких-либо когнитивных функций. И выдумывать отсутствие ментальности [у «других»] – это не довод, если учесть, например, что исходным допущением «когнитивных наук» предполагается реальность и познаваемость ментального подобно тому, как в физике предполагается реальность и познаваемость физических объектов.

6. Футурологический отзыв (Беркли)

«Вся ваша аргументация основана на той предпосылке, что ИИ – прерогатива только аналоговых и цифровых компьютеров. Однако эти компьютеры – лишь сегодняшний этап развития технологии. Чем бы не являлись ваши причинные зависимости, которые, как вы утверждаете, необходимы для интенциональности, со временем будут созданы такие устройства, которые станут обладать данными причинными зависимостями. Такие устройства, как и сегодняшние компьютеры – это тоже искусственный интеллект. Поэтому ваши аргументы никоим образом не направлены на возможность ИИ [по крайней мере, в будущем,] искусственно продуцировать когнитивные способности и объяснять их».

На этот отзыв у меня нет ответа. Только одно возражение: отзыв низводит проект сильного ИИ до банального уровня, определяя ИИ как то, что «искусственно продуцирует когнитивные способности и объясняет их». Интересно то, что данное утверждение по сути является выдвинутым от имени ИИ точным, чётко определённым тезисом о том, что ментальные процессы – суть вычислительные процедуры над формально определёнными элементами. Меня удивила подмена понятий. Если утверждение переопределяется на протяжении рассуждения, то мои возражения становятся бессмысленными, так как данный тезис – недоступная для проверки гипотеза.

Теперь настало время вернуться к вопросу, на который я обещал ответить: пусть по условиям эксперимента я понимаю английский язык и не понимаю китайский, а машина не понимает ни по-английски, ни по-китайски. То есть во мне должно существовать нечто, способствующее пониманию английских текстов, и отсутствие чего вызывает непонимание китайских текстов. Почему же невозможно придать это «нечто» (неважно, что) машине?

Я не вижу принципиальных причин, которые запретили бы приписывать машине способность понимать английский или китайский языки. Ведь в значительной степени человеческое тело и мозг - тоже машина. Однако я осознаю и убедительность аргументов, утверждающих о невозможности создания машины, способной понимать, если способ функционирования машины определяется исключительно в терминах вычислительных операций над формально определёнными элементами, то есть задаётся в виде компьютерной программы. Я понимаю английский язык и обладаю другими формами интенциональности не в силу того, что являюсь реализацией компьютерной программы (допустим, я собственно, есть реализация компьютерной программы). Нет, моя интенциональность возможна в силу иных причин: я – организм определённого вида, с определённой биологической (т.е. химической и физической) структурой, и данная структура при определённых обстоятельствах способна продуцировать причинно зависимые от неё интенциональные феномены (перцепции, действия, понимание, обучение и др.) И одним из аспектов моего аргумента является то, что обладать интенциональностью может только нечто, включающее в себя подобного рода причинные зависимости. Возможно, что физические и химические процессы, протекающие в существах, отличных от человека, могут приводить в точности к тем же [интенциональным] эффектам. Например, марсиане также могут обладать интенциональностью, но при этом их мозг может состоять из вещества, отличающегося от нашего. Всё это – вопросы эмпирического плана. Они сродни вопросу, возможен ли фотосинтез посредством химических процессов, в которых не принимает участия хлорофилл. Основной же пункт моего аргумента – это [теоретическое] утверждение: ни одна чисто формальная модель не самодостаточна для интенциональности, так как формальные свойства сами по себе не конституируют интенциональность. Они самостоятельно не обусловливают никаких возможностей, помимо возможности порождения следующего уровня формализации. Но и данная возможность осуществима при условии того, что запущена и работает машина. Все же случайные зависимости, сопутствующие конкретной реализации формальной модели, не существенны – ведь всегда можно определённую формальную модель положить в основу иной реализации, в которой случайные зависимости будут отсутствовать. Даже если каким-то чудом говорящие на китайском языке вдруг станут реализациями программы Шэнка и данная программа будет внедрена в говорящих на английском языке, в водопроводные трубы или в компьютеры, то всё равно ни трубы, ни компьютеры не станут понимать по-китайски.

Функционирование мозга – это вовсе не формальная тень, отбрасываемая последовательностью синапсов, а реальные свойства этих последовательностей. Все аргументы сильной версии искусственного интеллекта – это образы, рисуемые по чёткому контуру теней, отбрасываемых когнитивной способностью, с последующими заявлениями, что тени на самом деле и есть искомая реальность.

В заключение хочется обсудить некоторые общие философские проблемы, связанные с моим аргументом. Для ясности представим обсуждение в форме вопросов-ответов и начнём со старого, избитого вопроса:

«Может ли машина мыслить?»

Определённо, да. Люди сами в точности такие [мыслящие] машины.

«Да, но способна ли мыслить искусственная машина, созданная человеком?»

Я полагаю, что можно создать искусственным путём машину с нервной системой, нейронами, аксонами, дендритами и всем остальным, в точности подобными человеческому. Тогда ответ на вопрос представляется мне весьма определённым: да. Если в точности дуплицировать причины, то можно дуплицировать и следствия. Более того, полагаю, что можно продуцировать сознание, интенциональность и всё остальное на основе химических законов, отличающихся от законов протекания химических реакций в человеке. Как я уже говорил, это - вопрос эмпирический.

«Хорошо, но способен ли мыслить цифровой компьютер?»

Если под «цифровым компьютером» понимается всё то, что может быть описано в форме реализации компьютерной программы, тогда ответ снова, конечно, да. Ведь нас самих можно представить в форме реализации некоторых компьютерных программ, а мы способны мыслить.

«Но может ли мыслить, понимать и т.п. *в действительности сам* компьютер, снабжённый некоторой программой соответствующего типа? Может ли реализация некоторой программы, конечно, правильной программы, быть самодостаточной для понимания?»

Я считаю это хорошим вопросом, хотя его обычно путают с выше приведёнными. Ответом на него будет – нет.

«Почему нет?»

Потому что манипулирование формальными символами само по себе не предполагает интенциональности, оно полностью бессмысленно; оно даже не является манипулированием *символами* – такие символы ничего не обозначают. В терминологии лингвистов это – только синтаксис, но не семантика. Интенциональность, которой якобы обладают компьютеры, на самом деле, имеется лишь в сознании у программистов и пользователей, у тех, кто вводит данные и интерпретирует выходные результаты.

Цель примера с китайской комнатой и состояла в том, чтобы это показать. Мы помещаем в систему нечто, на самом деле обладающее интенциональностью (человека). Его действия программируются и мы убеждаемся в отсутствии у формальной программы, задающей его поведение, какой-либо дополнительной интенциональности. Программа, к примеру, не

добавляет ничего к человеческой способности понимать китайские тексты.

Именно данная черта ИИ – казавшееся столь привлекательным различие между программой и реализацией – оказывается фатальной для утверждения, что модель может быть дубликатом. Различие между программой и её реализациями на различной технической базе кажется аналогичным различию между ментальными и мозговыми операциями. И если бы можно было описать ментальные операции в виде формальной программы, тогда, вероятно, можно было бы описать сам разум, не прибегая ни к интроспективной психологии, ни к нейрофизиологии мозга. Однако уравнение «разум относится к мозгу так же, как программа относится к техническим средствам» неверно по ряду причин. Среди них выделим три основных:

Во-первых, из различия между программой и её реализацией следует, что одна и та же программа может иметь самые бестолковые реализации, в которых совершенно невозможно увидеть каких-либо проявлений интенциональности. Например, Вейзенбаум (1976, гл. 2) приводит детальное руководство по созданию компьютера с использованием рулона туалетной бумаги и кучки каменей. Схожим образом программа понимания китайского текста может быть представлена в виде серии водопроводных труб, набора ветряных мельниц или действий одноязычного носителя английского языка. Однако такое программирование не способствует пониманию китайского языка. Прежде всего, камни, туалетная бумага, ветер и водопроводные трубы – это не тот материал, который может обладать интенциональностью. Ею может обладать только то, что имеет такие же причинные зависимости, которые присущи человеческому мозгу. Но и этого недостаточно – хотя носитель английского языка и представляется наиболее подходящим материалом для интенциональности, мы видели, что интенциональности в нём не прибавляется ни за счёт запоминания ни за счёт использования программы. Знание программы китайскому не научает.

Во-вторых, программа – это чисто формальная конструкция. Интенциональные же переживания невозможно представить исключительно формальным способом. Они представимы в терминах содержания, но не формы. К примеру, убеждение в том, что идёт дождь, определяется не с помощью каких-то формальных средств, а посредством определённого психического состояния, которому присущи внимание, оценка и т.п. (см. Серль, 1979). Для убеждения, подобного тому, что идёт дождь, даже невозможно подыскать определённой синтаксической формы – оно может быть передано посредством неопределённого числа различных синтаксических выражений в различных языковых системах.

В-третьих, как ранее указывалось, ментальные состояния и события – это, буквально, продукты деятельности мозга. Но программа в этом смысле не является продукцией компьютера.

«Но если программы никак не конституируют ментальные процессы, почему так много людей убеждены в обратном?»

Данный вопрос требует разъяснений. Сам я не нахожу ответа. Идея того, что компьютерные модели способны вторгнуться в реальность и заменить её, с самого начала должна показаться подозрительной. Область применения компьютера не ограничивается моделированием ментальных операций. Никому ведь не приходит в голову, что компьютерная модель пожарной тревоги спалит окрестности, а компьютерная модель грозы промочит до нитки. Так с какой стати кому-то приходит в голову, что компьютерная модель понимания действительно что-то понимает? Некоторые утверждают о трудностях построения компьютеров, которые смогли бы почувствовать боль или влюбиться. Но и боль, и любовь нисколько не сложнее и не проще, чем понимание. Всё, что требуется для понимания, это подходящие входные и выходные данные, а также программа между входом и выходом, преобразующая первое во второе. И это, по сути, всё, что требуется компьютеру для выполнения любой задачи. Путать симуляцию с дупликацией, модель с дубликатом – такая же ошибка, идёт ли речь о понимании, боли, любви, пожаре или грозе.

Всё же есть ряд причин, в силу которых некоторым кажется, что ИИ в определённом смысле способен воспроизводить и таким образом объяснять ментальные феномены. Вряд ли удастся ликвидировать подобные заблуждения до тех пор, пока полностью не будут выявлены эти причины.

Первая и возможно самая важная причина заключена в путанице с понятием «обработка информации». Многие, имеющие дело с когнитивной наукой, убеждены в том, что человеческий мозг и его разум осуществляют то, что называется «обработкой информации». Аналогично, компьютер и его программа осуществляют обработку информации. Пожары и грозы, с другой стороны, информацию не обрабатывают. Таким образом, хотя компьютер и может моделировать формальные признаки любого процесса [например, пожара и грозы], следует выделять особые отношения, в которых компьютер находится с разумом и мозгом. Если компьютер должным образом запрограммирован, в идеале той же самой программой, под управлением которой функционирует мозг, то обработка информации и в компьютере, и в мозгу протекает идентичным способом. Такая обработка информации и составляет суть ментального. Проблема данного аргумента состоит в том, что он опирается на неясность понятия «информация». Запрограммированный компьютер информацию не обрабатывает в том смысле, в котором люди «обрабатывают информацию» во время размышления над арифметической задачей или при чтении текстов и поиска ответов на вопросы об их содержании. Всё, что осуществляет компьютер, это манипулирование формальными символами. Способ обозначения объектов реального мира, который используют программист и пользователь про-

граммы, находится далеко вне зоны компетенции компьютера. Повторюсь, компьютер работает с синтаксисом, но не с семантикой. Если в компьютер ввести «2 плюс 2 равно ?», он выдаст «4». Но он абсолютно не понимает, что «4» означает 4 или вообще что-либо обозначает. И дело не в том, что ему не хватает какой-либо вторичной информации об интерпретации первичных символов, а в том, что первичные символы для него никак не интерпретируемы. Всё, что имеется у компьютера, это множество формальных символов. Таким образом, введение понятия «обработка информации» влечёт за собой дилемму: либо мы включаем в понятие «обработка информации» понятие интенциональности, либо мы этого не делаем. В первом случае компьютер не осуществляет обработку информации, а всего лишь манипулирует формальными символами. Во втором случае, хотя компьютер и осуществляет обработку информации, однако делает он это в том же смысле, в каком осуществляют «обработку информации» счётные и пишущие машины, желудки, термостаты, грозы и ураганы, а именно: в способе описания их характеристик выделяется информация на входе, информация на выходе и способ преобразования первой во вторую. И тогда стороннему наблюдателю приходится интерпретировать вход и выход как информацию (в обычном понимании этого слова). В связи с этим, во втором случае невозможны аналогии между компьютером и мозгом в терминах сходства процессов обработки информации.

Вторая причина – то, что ИИ присущи проблемы, связанные с уже устаревшими концепциями бихевиоризма и операционализма. Поскольку структура входных-выходных данных у должным образом запрограммированных компьютеров может быть схожей со структурой входных-выходных данных у людей, мы склонны объявлять ментальные состояния у компьютера сходными с ментальными состояниями людей. Однако следует преодолеть этот соблазн, увидев, что и теоретически, и эмпирически система может обладать человеческими способности и при этом не обладать интенциональностью. Моя настольная счётная машина обладает вычислительными возможностями, но интенциональность ей не присуща. В данной работе я пытался показать, что если система способна работать со входом и выходом, копируя способности китайца, то всё равно вне зависимости от используемой программы система не будет понимать китайские тексты. Тест Тьюринга – типичный представитель беззастенчивой бихевиористской и операционалистской традиций. Я уверен, что если исследователи ИИ окончательно отвергнут бихевиоризм и операционализм, то исчезнет и большая часть путаницы между моделью и дубликатом.

Третья причина в том, что уже давно отживающий операционализм связывается с умирающей формой дуализма. В самом деле, сильный искусственный интеллект обладает правом на жизнь только на основе дуалистического допущения: там, где излишне сильно пекутся о разуме, мозгу не место. В сильном искусственном интеллекте (равно как и в функционализме) первостепенное значение уделяется программам. Они независимы от способов реализации в машинах. И действительно, одна и та же программа может быть реализована и в электронной машине, и в декартовой «мыслящей субстанции» и в гегелевском «абсолютном духе». Наиболее удивительным и, пожалуй, единственным моим открытием в ходе работы с данным материалом, явилось то, что многих исследователей ИИ шокировала сама идея о зависимости реального феномена человеческого мышления от реальных физико-химических свойств реального человеческого мозга. Но если немного задуматься об этом факте, то можно понять, что, собственно, удивляться и нечему. Ибо пока вы не примите одну из форм дуализма, у проекта сильного ИИ нет шансов. Проект заключается в том, чтобы посредством программ воспроизвести и объяснить ментальное. Однако его завершить не удастся, пока разум не только теоретически, но и эмпирически независим от мозга, ибо программа совершенно независима от любой реализации.

Пока вы не поверите, что теоретически и эмпирически разум отделим от мозга (это – дуализм в сильной форме), вы не можете надеяться воспроизвести ментальное путём написания и запуска программ, поскольку программы должны быть независимы от мозга и иных частных форм реализаций. Если ментальные операции суть вычислительные операции над формальными символами, то, следовательно, у них нет сколько-нибудь значимой связи с мозгом. Единственной связью может оказаться то, что мозг – это машина, которая, наряду с другими машинами, разнообразие которых бесконечно, способна к реализации программ. Данная форма дуализма не является традиционной декартовой раздвоенностью, утверждающей существование двух типов *субстанций*. Однако она соглашается с ней в том, что настаивает на отсутствии существенной связи между специфически ментальным в разуме и материальными свойствами мозга.

Дуалистическая основа скрывается от нас наличием в литературе по проблеме ИИ частых выпадов против дуализма; о чём авторы, кажется, не осведомлены, так это о том, что сама их позиция основывается на сильной версии дуализма. «Может ли машина мыслить?» Моя собственная точка зрения заключается в том, что *только* машина и может мыслить, причём лишь особенные виды машин, такие, как мозг, или машины, обладающие причинными зависимостями, подобными причинным зависимостям мозга. Также я уверен в том, что сильный ИИ ничего вразумительного не может сказать о мышлении по причине того, что он ничего не может нам рассказать о машинах. По его собственному определению, дело в программах, а программы не есть машины. Чем бы интенциональность не являлась, по сути своей она биологический феномен и находится в причинной зависимости от таких специфических биохимических источников, как лактация, фотосинтез или любой другой биологический феномен. Никому не придёт

в голову, что мы можем получать молоко и сахар, запуская компьютерную модель формализованных последовательностей лактации и фотосинтеза. Тем не менее, в отношении разума многие люди готовы поверить в чудо по причине глубоко укоренившегося, извечного дуализма: разум, по их предположениям, является разновидностью формальных процессов и не зависит от весьма специфических материальных причин в том смысле, в котором молоко и сахар от них зависят. В защиту этого дуализма часто выставляется довод, что мозг является цифровым компьютером (ранние компьютеры, кстати, так и назывались, «электронным мозгом»). Но это не помогает. Конечно, мозг есть цифровой компьютер. Поскольку всё вокруг – это цифровой компьютер, то и мозг тоже. Суть в том, что мозговая причинно-зависимая способность продуцировать интенциональность не может заключаться в реализации компьютерной программы, так как, в принципе, можно создать любую программу на любой вкус, однако при этом психических способностей [у компьютера] не прибавится. В чём бы ни заключалась деятельность мозга по обеспечению интенциональности, она не может основываться на реализации программы, потому что сама по себе программа для интенциональности не достаточна [3].

[3] Я в долгу перед очень большим числом людей за возможность обсуждения данной темы и за их терпеливые старания помочь преодолеть моё невежество в отношении искусственного интеллекта. Особенно я хотел бы поблагодарить Нэда Блока, Хьюберта Дрейфуса, Джона Хогленда, Роджера Шэнка, Роберта Виленски и Терри Винограда (*прим. авт.*)

Психологизм и бихевиоризм

Нед Блок

Перевод Ирины Матанцевой и Ивана Чижова

Перевод выполнен по работе**: Block, N. 1981, Psychologism and Behaviorism, Philosophical Review 90, pp. 5–43.**)

В данной работе я буду рассматривать психологизм как руководство по различению интеллектуального поведения от неинтеллектуального. Согласно психологизму, поведение системы задаётся внутренними параметрами информационного процесса. Внешне поведение систем может быть *неотличимо* от интеллектуального поведения и абсолютно тождественно ему как в актуальном, так и в потенциальном отношении. Поведенческие диспозиции и корреляты – ничем не отличаются. Однако при этом одна система - интеллектуальная, другая – нет, что обусловливается различиями в информационных процессах, протекающих между стимулом и реакцией.

Основными положениями данной работы является следующее. Во-первых, психологизм правомочно использовать для анализа интеллектуального поведения и его ни в коем случае нельзя игнорировать, как это принято в бихевиоризме. Во-вторых, стандартные аргументы против бихевиоризма недостаточны ни для опровержения бихевиористской концепции мышления, ни для укрепления позиций психологизма.

Безусловно, психологизм – проклятие для бихевиористов[1]. Однако и многие философы, которые отнюдь не относят себя к лагерю бихевиористов, заблуждаются относительно психологизма. Например, Майкл Дамит утверждает:

> «Если бы марсианин научился говорить на человеческом языке, или если бы был разработан робот, поведение которого не отличалось бы от поведения носителя языка, тогда и марсианину и роботу можно было бы приписать знание правильного способа означивания языковых выражений с таким же правом, как оно приписывается и носителю языка, несмотря на то, что при этом внутрен-

[1] На самом деле Райл в «*Понятие сознания*» (Ryle, 1949) открыто нападает на психологизм. Он считает, что именно мы являемся судящими ``в суждении, что кто-то интеллектуален или неинтеллектуален". Далее он заключает: ``Наше исследование направлено не на причину, а на мастерство, привычку, обязательство и склонность." См. работу Джерри Фодора *Психологическое разъяснение* (Fodor, 1968), где он проницательно критикует позицию Райла по поводу психологизма.

ние языковые механизмы совершенно различаются». (Dummett, 1976)

На первый взгляд, позиция Даммита связывает ментальные состояния с актуальным или потенциальным поведением. Для такой связи внутренний информационный процесс становится *не* принципиален – он лишь аффектирует поведение, оказывает на поведение некоторое неопределённое воздействие. В суждении Даммита, однако, можно обнаружить крупицу правды. Эту толику правды, проясняющую позицию Даммита, многие философы ошибочно полагают несовместимой с психологизмом.

Предположим, марсиане вошли в контакт с нами. При этом оказалось, что поведение марсиан ничем не отличается от поведения людей. Мы учим их язык, они учат наш. Между нами развиваются обширные деловые и культурные связи. Мы печатаемся в их журналах, они – в наших. Мы смотрим их фильмы, они – наши. Через некоторое время психологи, как с той, так и с другой стороны, начинают исследовать психологические механизмы людей и марсиан. Исследования вскрывают различия в самих основаниях этих механизмов. Психологи пробуют объяснить – почему при различиях в психологических механизмах поведение марсиан не отличается от поведения людей. Возможно, объяснение будет исходить из предположения, что и люди и марсиане – суть продукты реализации некого плана творения. Данному плану творения, как и любому проекту, например, проекту искусственного интеллекта, может быть присущ ряд технических альтернатив. Допустим, в соответствии с этим планом создается машина, которая рассуждает на английском языке. Тогда одна альтернатива плана творения может состоять в репрезентации информации в машине на английском языке и в формулировке правил вывода на этом же языке. Другой подход может состоять в создании искусственного языка, на котором выражаются логические формы предложений и в построении транслятора с английского на этот искусственный язык. Второй подход упрощает правила логического вывода, но увеличивает вычислительные затраты для реализации процедуры перевода. Теперь предположим, что и марсианские и человеческие психологи сошлись во мнении о том, что различия марсиан и людей обусловлены различиями в альтернативах плана творения. Должны ли мы заключить отсюда, что марсиане *не* разумны? Конечно, нет! Это было бы грубым человеческим шовинизмом. Я полагаю, что многим философам не нравится психологизм в силу того, что они (ошибочно) считают его проявлением человеческого [интеллектуального] шовинизма.

Одна из целей данной статьи состоит в раскрытии того положения, что психологизм совершенно не предполагает шовинизм.

Показав правоту психологизма, я помогу и приободрю тех, кто сомневается в функционализме (точке зрения, согласно которой ментальные состояния – суть функциональные состояния, т.е. состояния, определимые

в терминах причинных ролей (casual roles). Их сомнения в некоторой степени предопределены верой функционалистов в возможность существ, которые выглядят и действуют, как люди (обладая при этом сетью внутренних состояний, причинные зависимости которых воспроизводят наши ментальные состояния). Такие существа отличаются от людей тем, что ими управляет сеть гомункулусов, имитирующая функциональную организацию личности. Вера в гомункулусов приводит к утверждению о том, что в «гомункулусной» голове ментальности нет. В ответ на это функционалисты «проглатывают пулю». Они утверждают примерно следующее: «Если я обнаружу, что у моего лучшего друга или у наиболее ценного сотрудника «гомункулусная» голова, то этот факт не принудит меня считать его неразумным (и не повлияет на мое мнение о других аспектах его психики). Для интеллекта различия во внутренних действиях не существенны, если эти различия не влияют на актуальное или потенциальное поведение (или поведенческие контрфакты)». Если в данной статье будет продемонстрирована истинность положений психологизма, то подобная стратегия защиты функционализма станет невозможной.

Главная линия нашего доказательства будет строится на основе анализа хорошо известного теста Тьюринга. Тест заключается в следующем. В одной комнате находится машина, в другой комнате – человек. В течение установленного периода времени, например, часа, и человек и машина отвечают на вопросы человека (судьи), который находится в третьей комнате. Машина проходит тест тогда, когда судья не может определить, какие ответы выдаёт машина, а какие – человек. Вначале данный тест отражал взгляды современников Тьюринга на то, что значит быть интеллектуальным. А именно, интеллектуальное – это то, что действует интеллектуальным способом. Определить понятие «интеллектуальный» крайне трудно, однако «интеллектуальность» легко распознать по соответствующему поведению.

Обратите внимание, что слово «интеллектуальный», приведённое выше, да и вообще в контексте обсуждения теста Тьюринга,[2] употребляется нами *не* в смысле того, когда говорится, что один человек умнее другого. «Интеллектуальность» означает то, что нечто обладает мыслью или способно рассуждать.

Предложение Тьюринга популярно представлять некоторой версией операционализма. «Быть интеллектуальным» - значит пройти тест Тьюринга (или, в стиле Карнапа: если системе задан тест Тьюринга, то она интеллектуальна тогда и только тогда, когда его проходит). Операционалист-

[2] Сам Тьюринг говорил, что вопрос о том, может ли машина *мыслить*, следует «заменить на» вопрос, может ли она пройти тест Тьюринга, но большинство обсуждений тесте Тьюринга имели отношение к *интеллектуальности* в большей мере, нежели к разуму. (Статья Тьюринга (1950) была названа ``Вычислительные машины и *интеллект*" [курсив мой – Н.Б.])

ский вариант тьюринговской концепции интеллекта имеет недостаток, присущий всем иным формам операционализма. Это - предположение о *непогрешимости* измерительного прибора (в нашем случае таким «прибором» выступает тест Тьюринга). Нелепо выяснять *действительно* ли интеллектуально устройство, прошедшее тест Тьюринга. Так же нелепо выяснять причину, по которой устройство тест не прошло, хотя при всём при этом устройство может быть на самом деле интеллектуальным.

Данной проблемы можно избежать, если перейти от жёсткой операционалистской формулировки к бихевиористской диспозициональной формулировке. Бихевиористский подход близок к операционалистскому, но при этом он более гибок. Интеллект идентифицируется не со способностью прохождения теста, а с поведенческой диспозицией [*намерением*] его пройти. Неудачное прохождение теста уже не рассматривается столь строго. Теперь правомочно спросить о системе, которая не прошла тест - *на самом ли деле* она собиралась его проходить? Помимо этого, факт прохождения теста перестаёт выступать *неопровержимым* доказательством намерения его пройти – например, успех может быть случайным.

Однако и бихевиористская формулировка является предметом серьёзных затруднений. Очевидная проблема состоит в определении степени доверия к судье-человеку. Судья может *присуживать* человеку - например, не желая ставить машину в один ряд с человеком, он вообще не будет ставить машине положительных оценок. При этом поводом для распознавания машины может послужить машинообразный стиль ответов, что наблюдательный судья, несомненно, отследит.

В принципе, данную проблему легко устранить. Для этого следует модифицировать тест Тьюринга. Судья может не выяснять, какие ответы принадлежат машине, а какие - человеку. Его цель, скажем, судить о том, что один или оба отвечающих настолько же интеллектуальны, насколько интеллектуален средний человек. Однако подобного рода модификация теста приводит к порочному кругу - «интеллектуальность» определяется посредством суждений судьи об «интеллектуальности». Модификация бесполезна даже если не учитывать порочный круг – ведь прежняя проблема приобретает лишь новую форму: возможно, шовинист-судья станет отрицать интеллект у действительно интеллектуальной системы, опять же подметив машинный стиль мышления.

Более существенно то, что человека-судью можно легко одурачить совершенно неинтеллектуальной машиной. Этот факт прекрасно демонстрируется простейшей программой (Боден, 1977, Вайзенбаум, 1966). Программа разработана Джозефом Вайзенбаумом и состоит всего из двух сотен операторов на Бейсике. Применяя небольшой набор простых стратегий, программа может превосходно имитировать психиатра. Основная техника строится относительно ключевых слов - «я», «ты», «похожий», «отец» и «все». Слова упорядочены. Так, порядок слова «отец» выше, нежели чем у слова «все». Если например, вы напечатаете: «Мой отец всех боится», машина задаст один из вопросов относительно слова «отец», например: «Что еще приходит вам на ум, когда вы думаете о своем отце?». Если вы введёте фразу: «Я знаю, что все смеются надо мной», то получите вопрос относительно слова «все», например: «Как вы думаете, кто в большей степени?». Программа обладает и иными способностями. Например, она заменяет слова «ты» на «я» и «мной» на «тобой». Если вы напечатаете: «Ты со мной не согласен?», она в ответ может спросить: «Почему ты думаешь, что я с тобой не согласен?». Также программа *запоминает* предложения, содержащие некоторые ключевые слова, такие как «мой». Если *текущий* ввод не содержит ключевых слов, но *ранее* вы употребляли фразу: «Мой молодой человек заставил меня сюда прийти», программа может проигнорировать ваше текущее высказывание. Вместо прямого вопроса на текущее высказывание, она может спросить: «Это как-то влияет на то, что ваш молодой человек заставил вас прийти ко мне?» Если все эти уловки оказываются недостаточными, то у программы есть ряд других вопросов типа: «Кто здесь психиатр? Ты или я?»

Данная система *абсолютно* не обладает интеллектом. Однако практика показывает, что в ходе коротких бесед она *поразительно* неплохо обманывает людей. Конечно, машина Вайзенбаума не может обманывать длительное время, особенно если человек задался целью исследовать интеллектуальные возможности машины. Но удивительный успех программы (секретарша Вайзенбаума просила его выйти из комнаты, чтобы поговорить с машиной наедине) позволяет судить, что благодаря человеческой доверчивости чуть более сложная программа (причем далеко не интеллектуальная) может ввести в заблуждение практически любого человека-судью. Склонность к обману со стороны подобных программ зависит от степени нашей доверчивости, опыта работы с машинами и многих других факторов. Поэтому кажется неправомерным принятие точки зрения на природу интеллекта или мышления в тесной связи с человеческими суждениями. Может ли от доверчивости судьи зависеть то, что машина *в действительности* мыслит или является интеллектуальной?

В итоге, человек-судья может оказаться пристрастным шовинистом в отрицании действительно интеллектуальных машин. Но он может и излишне свободно признавать мыслящими те машины, которые всего лишь умело сконструированы.

Описанные выше проблемы разрешаются, если мы сможем чётко определиться с тем, что же именно в последовательности реакций на вербальную стимуляцию есть продукт мышления – продукт мышления в любых формах его проявления – как известного, так и неизвестного. Выявленный объективный фактор позволит устранить зависимость от субъективности мнения [судьи] в решении обозначенной в предыдущем абзаце проблемы. Допустим, что с позиции бихевиоризма *можно* неоспоримо оп-

ределить то, что же именно в последовательности вербальных выходных данных является, как мы скажем, «осмысляющим» последовательность входных данных. Конечно, понятие «осмысления» отнюдь не бесспорно. Мы *предполагаем* его очевидность и [возможность появления «смысла» в последовательности выходных сигналов]. Но даже такое предположение не спасёт концепцию интеллекта, предложенную в форме теста Тьюринга. В этом мы далее убедимся.

В тьюринговом определении интеллекта судья играет большую роль. Конституируя интеллектуальность, именно судья решает проблему фактической спецификации интеллектуального поведения или поведенческого намерения к мышлению. Отсюда можно предположить, что «осмысленность» (а это понятие, как мы договорились, можно определить очевидным образом), равносильно игнорированию одного из традиционных доводов против бихевиористов. Данный довод – невозможность бихевиористами специфицировать поведенческие диспозиции способом, не вызывающим сомнения. Конечно, это - огромная уступка бихевиоризму. Но в последующем она не будет играть значимой роли.

Настало время предложить нашу версию тьюринговой концепции интеллекта. Эта версия избегает описанных выше проблем:

> интеллект (или, более точно, диалоговый интеллект [, т.е. интеллект, проявляющийся в общении]) – это диспозиция [или предрасположенность] продуцировать осмысленную последовательность вербальных реакций на последовательность любых вербальных стимулов.

Выражение «любых» подчеркивает стремление избежать различного рода каверзных вопросов. Чтобы быть интеллектуальной, в соответствии с нашей концепцией, необходимо наличие стремления (намерения) у системы осмысленно отвечать не только на то, что собеседник *действительно* говорит, но также и на то, что он *мог бы* сказать.

Хотя новая формулировка существенно изменяет тьюринговую дефиницию (при нашем предположении возможности адекватного описания понятия «осмысленный»), все же эта формулировка несомненно, выполнена в бихевиоральном стиле. Представляет интерес рассмотрение стандартных аргументов против бихевиоризма – способны ли они опровергнуть концепцию интеллекта, сформулированную в виде теста Тьюринга?

Возможно, наиболее существенным доводом против бихевиоризма мы обязаны Чизхолму и Гичу (Chisholm, 1957). Предположим, бихевиорист анализирует [ментальное состояние] «желание мороженого». Анализ осуществляется путем представления этого желания в виде совокупности поведенческих диспозиций. К таким диспозициям относится, например, намерение человека схватить мороженое, особенно когда его «дарят». Но человек, желающий мороженое, намерен схватить его, если *знает*, что схватываемый предмет – именно мороженое (а не тюбик дегтя, например,

который предлагается шутки ради и который может иметь, как и мороженое, конусообразную форму). Далее, человек должен быть *убежден*, что принятие мороженого не влечет конфликта с другими *желаниями*, которые могут иметь для него большее значение (например, он избегает обязанности оказывать ответную услугу за мороженое). Короче, поведенческие диспозиции желания обусловлены всеми *остальными* ментальными состояниями желающего. Подобным образом следует осуществлять бихевиоральный анализ, например, убеждений и многих иных ментальных состояний. Вывод: невозможно определить условия поведенческой диспозиции некоторого конкретного ментального состояния без учета *других ментальных состояний*.

Другой традиционный аргумент против бихевиоризма вытекает из рассмотренного довода Чизхолма-Гича. Если индивидуальные поведенческие диспозиции зависят от некоторой совокупности ментальных состояний, то могут ли *различные* ментальные совокупности продуцировать одинаковые поведенческие диспозиции? Этот вопрос послужил началом серии контрпримеров, которые в целом называются аргументом «совершенный актер». Патнэм (1979b) приводит убедительные доводы: вообразите общество совершенных актеров (супер-спартанцы Патнэма). В силу установленных законов (законов спартанского общества) им нельзя выказывать боль, даже если они на самом деле её испытывают. В таких условиях никакую поведенческую диспозицию этих актеров бихевиорист не может связать с болью. Для определения факта боли необходимы выражающие боль поведенческие диспозиции [, а для супер-спартанцев их выявить невозможно]. Напротив, человек может притворяться, что ему больно. В последнем случае [, а именно его и называют «совершенным актером»] представляется недостаточным осуществление анализа только поведенческой диспозиции боли.

Другие традиционные контрпримеры, опровергающие бихевиоризм, уже менее значимы. Их демонстрируют аргументы «паралитики» и «мозги в бочке». Паралитики – это «обращённые» супер-спартанцы Патнэма. Они также испытывают боль без каких либо её внешних проявлений.

В дальнейшем, под «традиционными возражениями бихевиоризму» я подразумеваю три типа возражений: Чизхолма-Гича, «совершенные актеры» и «паралитики»[3].

---

3 Хотя возражение Чизхолма-Гича и возражение совершенных актеров, на мой взгляд, считаются главными возражениями бихевиоризму, у некоторых авторов они даже не входят в список возражений. Например, у Рорти (Rorty, 1979, p. 98) они не включены в данный список. Такие, как Рорти, придают большое значение общему возражению, которое я проигнорировал, а именно, что бихевиористский анализ психических состояний предполагает аналитическую истинность значений ментальных терминов. Я частично игнорировал аналитические возражения, потому что основные соперники бихевиоризма - физикалисты и функционалисты - обычно придерживаются версий, ко-

1. Опровергают ли бихевиористскую концепцию интеллекта стандартные возражения правомочности бихевиористской диспозиции?

Рассмотренные выше аргументы совершенно справедливо считаются основными опровержениями [правомочности] бихевиористского анализа психических состояний, например, анализа веры, желания, боли. Более того, эти аргументы привлекаются для демонстрации ложности бихевиористского анализа интеллекта. Согласно бихевиоризму интеллект вполне убедительно рассматривается как вторичное ментальное свойство. Первичными психическими состояниями выступают вера, желания и т.п. Они являются причиной изменений других, [менее существенных,] психических состояний. Так как аргументы разрушают бихевиористский подход к анализу первичных психических состояний, а интеллект – лишь производное от этих состояний, то можно смело заключить, что кажущийся убедительным взгляд бихевиористов на интеллект на самом деле не убедителен[4].

Однако было бы нечестно по отношению к бихевиоризму опровергать его таким простым способом. Бихевиористы, в общем, считают, что [поведенческие] диспозиции суть «чистые диспозиции». Например, согласно Райлу (Ryle, 1949) подчеркивается, что «обладать диспозициональным состоянием не означает находиться в каком-то конкретном состоянии или испытывать какие-то конкретные изменения». Так, например, уязвимость – это не причина поражения, а лишь *обстоятельство* лёгкости поражения. Подобным образом, приписать боль индивиду – это не значит раскрыть причинно-следственные зависимости состояния боли, а значит раскрыть, *что бы индивид делал,* когда ему было бы на самом деле больно. Однако только что упомянутое бихевиористское определение интеллекта как вторичного свойства выглядит правдоподобным тогда, когда первичные ментальные состояния понимаются как сущности, *обладающие причинными зависимостями*. Так как чистые диспозиции не обладают причинными зависимостями в прямом смысле понятия причинности, то анализ интеллекта как вторичного свойства должен показаться бихевиористу неубедительным, будь это самый правильный анализ интеллекта. Возможно, этим объясняется, почему бихевиористы и их сторонники выглядят так, как будто и не считают интеллект второстепенным свойством.

торые касаются приверженности аналитической истине (например, версия Льюиса и Шомейкера). Многие бихевиористы согласились бы, что логические связи, основанные на мысленном эксперименте, «слабее» аналитических. И я не вижу смысла исследовать сильные версии, т.к. бихевиоризм можно опровергнуть вне зависимости от аналитических результатов, а только путем приведения «слабых» доводов.

4 Здесь я в долгу у Сиднея Шумейкера.

Во-вторых, для бихевиориста вполне естественным представляется анализ интеллекта в соответствии с обозначенной выше линией, которая была названа мною как концепция интеллекта, основанная на тесте Тьюринга [ТТ-концепция интеллекта]. На бихевиоризм накладывается заплата, сшитая из широко известной операционалистской формулировки интеллекта. Не удивительно, что ТТ-концепция интеллекта популярна среди исследователей искусственного интеллекта[5]. ТТ-концепции интеллекта также симпатизируют многие философы, принимающие стандартные аргументы против бихевиористского анализа убеждений, желаний и пр.

Но в большей степени ТТ-концепция интеллекта привлекательна для бихевиористов тем, что *позволяет избежать традиционных возражений бихевиоризму*. И если это так, то просто глупо со стороны критиков бихевиоризма считать бихевиористский анализ интеллекта уничтоженным традиционными возражениями бихевиоризму, игнорируя при этом критику ТТ-концепции интеллекта. Поэтому следует изучить вопрос - насколько хорошо справляется ТТ-концепция интеллекта со стандартными возражениями?

Считается, что ТТ-концепция интеллекта предлагает необходимый и достаточный критерий интеллектуальности [системы]. На самом деле, традиционные возражения эффективно работают против необходимого условия. Для выработки достаточного критерия этих возражений недостаточно. Рассмотрим, например, аргумент совершенных актеров Патнэма. Суперспартанцы испытывают боль, хотя в поведении признаки боли не выражаются. Точно также машина может быть интеллектуальной, но при это не обладать намерением интеллектуально *действовать*, потому что, например, она может быть запрограммирована предполагать, что действовать разумно не *в ее интересах*. А как насчет противоположного случая? Совершенный актер, который притворяется, что *испытывает* боль, играет столь же правдоподобно, как супер-спартанец, притворяющийся, что *не чувствует* боли. Однако для того, чтобы играть в такого рода игру, иинтеллекта, по всей видимости, *не* требуется.

Теперь зададим вопрос – каким образом неинтеллектуальная система может в совершенстве притворяться интеллектуальной? По всей видимости, система, которая *так* хорошо притворяется интеллектуальной, на самом деле и должна быть интеллектуальной. Следовательно, невозможно привести поведенческую диспозицию, которая будет выступать необходимым условием интеллектуальности. Только лишь *на период обсуждения стандартного возражения* поведенческая диспозиция может служить достаточным условием интеллектуальности.

5 См. Schank and Abelson (1977). См. так же работы Weizenbaum's (1976) – критикующие программу Вейзенбаума (ELIZA).

Те же соображения можно привести и к возражению Чизхолма-Гича, которое утверждает, что диспозиция «болевого» поведения не является достаточным условием действительного обладания болью, так как диспозиция может задаваться рядом различных комбинаций психических состояний, например: {боль+обычная функция предпочтения [то есть обычное отношение к боли]} или {нет боли+желание обмануть, что боль есть}.

Обращаясь к интеллектуальному поведению, мы видим, что оно, как правило, продуцируется следующей комбинацией: {интеллектуальность+ обычная функция предпочтения}. Но может ли интеллектуальное поведение быть результатом комбинации {отсутствие интеллекта+желание обмануть, что интеллект есть}? В самом деле, возможна ли некая комбинация ментальных состояний и свойств, в которой *отсутствует интеллектуальность*, но которая вполне закономерно и бескомпромиссно продуцирует диспозицию интеллектуального действия? По всей видимости, такая комбинация невозможна. Поэтому возражение Чизхолма-Гича не опровергает утверждение ТТ-конепции интеллекта о том, что определенная диспозиция [интеллектуального поведения] достаточна для интеллектуальности.

И, наконец, примеры «паралитики» и «мозги в бочке» лишь более ограниченная [, по сравнению с предыдущими возражениям,] попытка применения возражений к необходимым - но отнюдь не к достаточным условиям психических состояний.

Только что обозначенный недостаток бихевиористского взгляда на интеллектуальность представляется умеренно серьезным, пока бихевиористы сосредотачиваются на достаточных условиях применения ментальных терминов (прежде всего, [например,] на связи между значением слова «боль» и обстановкой, в рамках которой мы обучаемся применять это слово).

Например, Тьюринг явно хотел сформулировать «достаточное условие» для своего бихевиорального, по сути, определения интеллекта[6]. Одна из моих целей в этой статье – исправить данный недостаток в стандартных возражениях бихевиоризму, показав, что ни одна поведенческая диспозиция не является достаточным критерием интеллектуальности.

6 Тьюринг утверждает: «Игра, возможно, может быть подвергнута критике на том основании, что случайности не имеют места в [детерминированной] машине. Если бы человеку пришлось притворяться машиной, он, естественно, показал бы довольно скудное поведение. Его бы выдала медлительность и неточность арифметических расчётов. Не могут ли машины произвести что-то, что обязательно представляется как продукт мышления и при этом данный продукт будет сильно отличаться от того, что делает человек? Это возражение очень сильно, но по меньшей мере, мы можем сказать, что если, тем не менее, машина может быть сконструирована так, что будет удовлетворительно играть в игру в имитацию, нам не нужно затрудняться по поводу этого возражения. (Turing, 1950, p. 435)

Только что я привёл довод в пользу того, что стандартные возражения бихевиоризму лишь частично эффективны против ТТ-концепции интеллекта. Продвинемся ещё на один шаг вперёд, таким образом перефразировав ТТ-концепцию интеллекта, что это *полностью* сведёт на «нет» силу стандартных возражений. Переформулировка состоит в замене в предыдущем определении слова «диспозиция» словом «способность».

Как было упомянуто ранее, существует многообразие причин, по которым интеллектуальная система не расположена действовать интеллектуально: убеждение в том, что действовать разумно не в её интересах, паралич и т.д. Тем не менее, интеллектуальные системы, которые не желают действовать разумно, или которые парализованы, все же имеют *способность* действовать разумно, даже если они не могут на практике применить эту способность.

Раскроем различия между бихевиоральной диспозицией [предрасположенностью, намерением] и бихевиоральной способностью.

Способность Θ не может проистекать из диспозиции Θ, пока не удовлетворены определенные *внутренние* условия – скажем, отсутствуют определённые взгляды и мотивы или в жилах не течёт кровь [, т.е. организм мёртв]. В самой простой форме диспозицию можно охарактеризовать множеством (возможно, бесконечным) условий ввода-вывода.

Если i1, то o1;

Если i2, то o2;

и так далее, [где «i» - вход (причина, условие и т.п.), «o» - выход (соответственно, следствие, состояние и т.п.)].

Первый удар по правомочности подобного рода спецификации бихевиоральной диспозиции наносится упоминанием *внутренних состояний* [они необходимы для раскрытия специфики отличия способности от диспозиции]:

Если sa и ia, то oa;

Если sb и ib, то ob;

и так далее,

где sa и sb – внутренние состояния. Для человека такими состояния будут, по крайней мере, вера, желания, чувства. Для машины внутренними состояниями будут, [конечно], ненментальные параметры, например, повреждённость предохранителей.

Приведённое выше различие между диспозицией [предрасположенностью] и способностью, имеет довольно поверхностный характер и нуждается в прояснении, особенно в отношении вопроса о том, какие типы внутренних состояний должны быть специфицированы в антецедентах. Если паралитики обладают бихевиоральными способностями, то описание внутренних состояний должно включать спецификацию функционирования механизмов ввода-вывода. Для системы, убежденной в том, что осу-

ществление интеллектуальных действий не в её интересах, спецификация внутренних состояний должна включать и описание того, что действовать интеллектуально – в ее интересах.

Заметим, однако, что бихевиористу не надо использовать *менталистских дескрипций* внутренних состояний – ему достаточно физиологических или функциональных описаний[7].

Читатель может засомневаться в том, что предложенная мною переформулировка бихевиоризма в терминах способностей помогает избежать стандартных возражений бихевиоризму лишь потому, что она *во многом* идёт на уступки. Ссылки на внутренние состояния – даже в психологическом или функциональном описаниях – могут показаться слишком большой уступкой психологизму (или другой небихевиористской теории). Отвечаю: чем больше [уступок], тем лучше для моих целей, ибо я намерен показать, что даже уступок *недостаточно* и путь от бихевиористских диспозиций к бихевиористским способностям *не спасает бихевиоризм*.

Настало время переформулировать [и обсудить] ТТ-концепцию интеллекта в свете предыдущих суждений. Назовем её *новой ТТ-концепцией интеллекта (нео-ТТ):*

> Интеллект (точнее, диалоговый интеллект) - это способность продуцировать осмысленную последовательность вербальных реакций на любую последовательность вербальных стимулов.

Кратко обсудим стандартные возражения бихевиоризму и покажем, что нео-ТТ их избегает. Во-первых, аргумент разумные паралитики и разумные «мозги-в-бочке» не обеспечивают контр-примеров, поскольку у них есть способность реагировать осмысленно, но нет средств реализации данной способности. Во-вторых, рассмотрим возражение «совершенных актеров». Очевидно, что разумное существо, в совершенстве имитирующее глупость, обладает способностью разумно реагировать. Кроме того, в отличие от случая с бихевиоральной диспозицией, уже никто не способен притворяться интеллектуальным, не являясь таковыми в действительности.

В-третьих, новая формулировка полностью обезоруживает возражение Чизхолма-Гича. Существует множество комбинаций верований и желаний, которые могут стать причиной того, что разумное существо будет не *предрасположено* давать осмысленные ответы. Однако эти верования и желания не разрушают *способности* существа давать такие ответы.

Более того, как я уже неоднократно упоминал, сложно представить, как некая совокупность психических состояний, не включающая интеллектуальных состояний, может привести к способности разумно реагировать на произвольную последовательность стимулов.

[7] Отступление от бихевиоризма, заключающееся в рассмотрении внутренних состояний, описанных физиологически или функционально смягчается. Физиологическое\функциональное описание при *правильном* анализе способностей может быть неопределенным в отношении сохранения бихевиористских особенностей теории.

Один завершающий момент. Заметьте, что моя уступка – это понятие «осмысленности». Данное понятие можно рассматривать в бихевиоральном контексте. Однако эта «уступка» не несет ответственности за тот факт, что нео–ТТ избегает стандартных возражений. На самом деле, последнее становится возможным в силу того, что, во-первых, трудно вообразить себе того, кто может быть интеллектуальным, на самом деле таковым не являясь и, во-вторых, в силу нашего перехода от понятия диспозиции к понятию способности.

2 Аргумент за психологизм и против бихевиоризма

Моя стратегия заключается в описании машины, которая производит (т.е. имеет способность производить) осмысленную последовательность вербальных реакций на вербальные стимулы. Машина является интеллектуальной в соответствии с нео-ТТ (и даже в соответствии с более грубыми версиями ТТ-концепции интеллекта).

В соответствии же с моей точкой зрения машина абсолютно неинтеллектуальна, в чём убеждают знания о том, какие внутренние информационные процессы в ней протекают.

Опишем нашу «неинтеллектуальную» машину. Введём термины. Назовем *печатаемой строкой* (строкой, для которой имеется возможность её напечатать) такую совокупность предложений, члены которых могут быть напечатаны последовательно друг за другом в течение часа или меньшего времени. Рассмотрим множество всех *печатаемых* строк. В английском языке это множество конечно, так как конечно множество слов (и поэтому число печатаемых строк ограничено). Множество печатаемых строк может быть очень большим, но все же конечным. Рассмотрим подмножество этого множества, которое содержит те и только те строки, которые самым естественным способом интерпретируются в форме диалога, и в котором хотя бы одна сторона что-то понимает при восприятии печатаемых строк. Назовем *осмысленной* строкой ту строку, которая может быть понята.

Предоставим каждому участнику диалога возможность «по очереди» высказывать по одному предложению (это допущение будет использоваться и далее). Если каждое из четных предложений в строке является разумным вкладом в разговор, то это – осмысленная строка. Не нужно строго определять, что именно считать осмысленным. Например, если предложение № 1 (нечётное) звучит так: «Давай посмотрим, как ты говоришь вздор», то предложение № 2 (чётное) будет осмысленно противоречивым.

Множество осмысленных строк, определенное таким образом, является конечным набором. В принципе, это множество полностью может быть создано усилиями громадного коллектива умнейших людей, которые работают длительное время, при [финансовой] поддержке внушительного

гранта и при помощи разнообразных технических средств. Люди, работающие в коллективе, *обладают воображением и принимают решение* по поводу того, какую строку считать осмысленной.

Возможно, программисты сочтут, что для создания действительно осмысленных строк им придется ставить себя на место определенных личностей с определенной историей. Они могут выбрать способ давать ответы, какие давала бы моя тетя Берта, которую ее рассеянный племянник привел в комнату с телеграфом и попросил время от времени отвечать на его глупые вопросы.

Представим совокупность осмысленных строк, записанных на ленту и используемых очень простой машиной следующим образом. Судья впечатывает предложение A. Машина перебирает список осмысленных строк, выбирая те предложения, которые начинаются с предложения A. Затем, она случайным образом выбирает одну из этих строк, начинающихся с A, и выводит второе предложение, назовем его B. Судья [в ответ на B] печатает предложение C. Машина просматривает список, выбирая те строки, которые начинаются с A, далее за которыми следует B, и далее – C. Машина выбирает одну из этих строк, начинающихся с ABC и печатает четвертое предложение, и так далее[8].

Возможно, читателю будет легче вообразить эту машину, если представить линейную последовательность предложений осмысленной строки в виде древовидной структуры. Предположим, первым начинает судья. Он печатает одно из предложений $A_1 \ldots A_n$. Программисты генерируют *один* осмысленный ответ на каждое из этих предложений, $B_1 \ldots B_n$. На каждое из $B_1 \ldots B_n$ судья может давать различные ответы и вопросы, так что под каждым из $B_1 \ldots B_n$ будет расти много ветвей. И вновь на каждый из этих ответов программисты генерируют некий один осмысленный ответ. И т.д.

Можно создать версию нашей машины. В новой версии все строки, начинающиеся с X, можно заменить единственным деревом с единственным символом X в вершине. Все строки, начинающиеся с XYZ, можно заменить ветвью этого дерева, в которой следующими вершинами будут Y и Z и так далее. Такая машина вместо поиска подстрок будет осуществлять поиск по дереву.

При условии правильного выполнения программистами своей работы, наша машина [вскоре] станет обладать способностью выдавать осмысленную последовательность вербальных реакций какими бы ни были при этом вербальные стимулы. Следовательно, согласно концепции нео-ТТ, машина является интеллектуальной. Однако *весь интеллект, который машина проявляет – суть интеллект программистов.*

8 Версия этой машины была обрисована в моей [статье] «Проблемы функционализма» (Block, 1978b).

Отметим также несущественность длительности теста Тьюринга, которую он установил равной одному часу. В принципе, машина, проходящая тест Тьюринга, может быть запрограммирована вышеуказанным способом для прохождения теста *любой* длины.

Отсюда я заключаю, что способности выдавать осмысленные ответы явно недостаточно для интеллектуальности. Таким образом опровергается нео-ТТ (вместе с предыдущей, более грубой, исходной ТТ-концепцией интеллекта и иными её модификациями). Я также заключаю, что критерий интеллектуальности частично обусловлен способом продуцирования интеллектуальной способности. Однако даже если поведение системы и актуально и потенциально не отличается от поведения разумного существа, систему нельзя считать интеллектуальной, если учесть, что протекающие в ней внутренние процессы похожи на внутренние процессы описанной машины. Это доказывает правоту психологизма [как руководства к определению интеллектуальности систем].

Я не раскрыл *полностью* то, о чём было вначале заявлено, поскольку не показал, что машина может дублировать способности отвечать [на вопросы], присущие реальной личности. Но и показанного вполне достаточно. Смысл примера очевиден для меня, особенно когда воображу, что программисты принимают решения *в точности* такие, какие делала бы тетя Берта. При этом основанием для принятия решения [об осмысленных строках] может служить психологическая или физиологическая теория тети Берты.

Теперь мы видим, почему психологизм не расходится с точкой зрения, высказанной ранее в связи с марсианским примером. Пример с марсианами посеял сомнения в существовании какого-то сингулярного информационного процесса, который вовлекается в производство всего интеллектуального поведения. (Я утверждал, что было бы шовинизмом подвергать сомнению разумность марсиан *только лишь* потому, что их внутренний информационный процесс сильно отличается от нашего информационного процесса).

Психологизм не шовинистичен. Он утверждает, что интеллектуальное поведение *не* есть продукт некоторого (по крайней мере, одного) внутреннего процесса определённого типа. Но настаивает на том, что поведение, имеющее определённые этиологические [показатели], не может считаться интеллектуальным без принятия допущения, что всё интеллектуальное поведение относится к тому же самому этиологическому «типу».

Суть примера с машиной можно лучше осветить, сравнив её с приемо-передающей радиоустановкой. Если разговаривать с разумной личностью (причём на любую тему) посредством радиоустановки, то мы будем слышать осмысленные ответы. Но это возможно отнюдь не в силу способности радиоустановки производить осмысленные ответы, которыми *она* располагает. Приемно-передающая радиоустановка – это моя машина, но

последняя действует [не как канал связи, а] как *канал передачи* интеллекта. Но при этом устройства различны, причём моя машина ключевым образом отличается от радиоустановки. В машине отсутствует непосредственная причинная зависимость ответов от вопросов судьи – вопросы не достигают тех, кто придумывает ответы. В случае радиоустановки, человек, придумывающий ответы, должен слышать вопросы. В случае же моей машины причинная действенность программистов ограничена – судья получит лишь то, что программисты вложили перед началом его опроса.

Читателю также следует обратить внимание на то, что мой пример является расширением традиционного довода «совершенных актеров» - машина «притворяется» интеллектуальной, не являясь таковой на самом деле. Теперь легко видеть, что личность может обладать способностью отвечать разумно, даже если интеллект, ею выказываемый, *не принадлежит ей* – например, если человек заучивает наизусть ответы на китайском, хотя понимает только английский[9]. Идиот с фотографической памятью, такой, как Luria's, известный мнемотехник, может вести блестящую философскую беседу, если он запомнил тексты великих философов.

Машина, которую я до сих пор описывал, ограничена тем, что на входе и выходе у неё – машинописные тексты. *Но это ограничение не существенно, мой аргумент значим и для полнокровной бихевиористской теории интеллекта,* а не только для теории диалогичного интеллекта. Остаётся это доказать. Для этого мне потребуется показать, что допущения, применяемые по поводу [вербальных стимулов и реакций, т.е.] машинописных текстов, также применимы и для чувственных стимулов и реакций. И если это так, то идея описанной мною машины сводится к идее робота, который интеллектом не обладает, однако в любой возможной ситуации действует столь же разумно, как и человек.

Для доказательства мне потребуются как эмпирические, так и теоретические подтверждения. Эмпирические подтверждения слишком долго детально раскрывать, поэтому они будут лишь кратко упомянуты. Во-первых, то, что мне известно из области физиологии чувств, уже вполне достаточно для подкрепления утверждения, что каждый чувственный стимул можно «квантовать» без каких-либо погрешностей при воздействии этого стимула на мозг или на поведение. Главное условие – чтобы величина «квантования» была достаточно малой.

Представим, что наши органы чувств от внешнего мира отделяет перегородка, предназначенная для аналого-цифрового преобразования. Например, если бы стимулирующие параметры были измеримы, то данная перегородка вместо них вырабатывала бы числа - если параметр имеет значение 0,111…, то на выходе перегородки получаются значения 0,11. При условии того, что величина «кванта» в достаточной степени мала (т.е.

9 Данная точка зрения более детально рассмотрена в конце статьи.

не слишком много десятичных разрядов отброшено), преобразование из аналоговой формы в цифровую не должно приводить к каким-то психическим или поведенческим отличиям в способах стимуляции без перегородки и с нею. И если это так, тогда любой может воспринимать результат аналого-цифрового преобразования как значимый стимул, и таким образом, в некотором ограниченном промежутке времени будет возможно лишь конечное число комбинаций стимулов.

Мне указывали, что подобный вывод правомочен при рассмотрении *любого* физического объекта в виде системы со входом и выходом. Здесь важно то, что ни одна физическая система не может быть бесконечно мощным усилителем, поэтому вполне ограниченная «мощность усиления» данной физической системы может таким образом производить квантование входных данных, что оно никак не влияет на выходные данные. Я недостаточно хорошо знаю физику поэтому далее не буду заниматься физическими подтверждениями моего аргумента.

Моя линия аргументации имеет в большей степени теоретический, нежели эмпирический характер. Идея состоит в том, что наше *понятие* об интеллектуальности позволяет снабдить разумное существо квантованными сенсорными устройствами. Предположим, например, что глаза марсиан – это видеокамеры, в которых информация, передаваемая в мозг, эквивалентна газетным, «точечным» картинкам. Т.е. глаза – это матрицы, содержащие большое количество ячеек, каждая из которых может быть либо белой, либо черной (марсиане не различают цвета). Если марсиане не только внешне поразительно похожи на нас (в поведении), но и в особенностях протекания внутреннего информационного процесса, то становится невозможным считать их глаза-камеры (и другие конечные органы чувств) признаком неразумности. Заметим, что существует лишь конечное число таких «точечных» образов, для которых фиксирована площадь точки. Поэтому существует и конечное число возможных *последовательностей* таких образов заданной длительности, и, следовательно, конечное число *возможных визуальных стимулов* заданной длительности.

Легко видеть, что и эмпирический и концептуальный подходы поддерживают утверждение, что интеллектуальное существо может иметь конечное число возможных последовательностей стимулов за конечное время (это верно и для реакций). Но тогда последовательности стимулов программисты могут систематизировать в каталоге таким же образом, как они систематизировали вопросы судьи в машине, описанной выше. Таким образом, робот, запрограммированный также, как запрограммирована описанная ранее машина, может обладать любой бихевиоральной способностью, которой обладает и человек. Следует подвести итог. До настоящего времени мы имели дело с поведенческой оценкой диалогового интеллекта. Сейчас мы расширили область действия аргумента и применили его к общей бихевиористской теории интеллекта. Однако это не дало принципи-

ально новых выводов. Поэтому далее, для удобства и простоты можно ограничиться обсуждением диалоговой интеллектуальности, сути раскрываемых положений это ограничение не затрагивает.

К этому времени у читателя, возможно, возник ряд возражений. Например, интенсивное использование фразы «в принципе» может привести к тому, что выражение «в принципе возможно» означает принципиальную возможность построения *любой* из описанных мною машин. А это, конечно, весьма странно. Читатель так же может возразить: «Способность вашей машины проходить тест Тьюринга не зависит от времени тестирования» или: «Вы просто вводите новое значение слова «интеллектуальность». Или, возможно, ему интересно будет узнать, как *я* сам смог бы себе возразить.

В дальнейшем я попытаюсь ответить на эти и другие возражения. Если возражение обозначено номером с индексом (например, 3а), то это значит, что оно относится к предыдущему возражению или ответу. Читатель может пропустить любое возражение или ответ на возражение без потери [содержательной] связности [статьи].

[3. Дискуссия]

Возражение 1.

Ваш аргумент слишком сильный в том смысле, что о любой интеллектуальной машине можно утверждать, что проявляемый ею интеллект суть интеллект программистов.

Ответ.

Я *не* утверждаю того, что *любая* машина, созданная разумными существами, является интеллектуальной исключительно благодаря интеллекту этих существ. Такой подход не прослеживается в моих доводах. Если мы когда-то и сделаем [настоящую] интеллектуальную машину, то непременно оснастим её механизмами самообучения, решения проблем и т.д. Возможно, будут открыты универсальные принципы обучаемости, общие принципы решения проблем и т.д. и эти принципы мы сможем встроить в машину. Но хотя мы и *создаем* машину интеллектуальной, проявляемый ею интеллект – это в той же мере её собственный интеллект, в какой мере наш разум – наш собственный разум, а он, в основном, является результатом чрезвычайных достижений наших предков.

Для контраста, если моя машина, отыскивающая строки, выдает хитроумную последовательность слов P в ответ на вопрос C, тогда последовательность CP, дословно, является той, которая была выдумана и заложена программистами. Возможно, программисты скажут об одном из своих коллег: «Джонс придумал такую игру слов – это он такой умный!».

Проблема с концепцией нео-ТТ (и её предшественниками) состоит в том, что она не позволяет находить отличия между поведением, которое отражает *собственный* интеллект машины и поведением, отражающим

только интеллект программистов, сконструировавшими машину. Я предлагал, что только частное этиологическое объяснение интеллектуального поведения способно решить эту проблему.

Возражение 2.

Если бы строки были записаны до начала текущего года, машина не смогла бы ответить, как человек, на предложение типа: «Что вы думаете о последних событиях на Ближнем Востоке?»

Ответ.

Систему можно считать интеллектуальной, даже если в неё не заложены знания о текущих событиях. Более того, машина может *имитировать* интеллектуальность без *имитации* знаний о текущих событиях. Программисты могли бы при желании построить модель разумного Робинзона Крузо, который ничего не знает о том, что в последние 25 лет произошло в мире. Также программисты могли бы взять на себя обязанность периодически перепрограммировать машину для того, чтобы она имитировала знания о текущих событиях.

Возражение 3.

Вы утверждали, что машина с определенной внутренней механической структурой не интеллектуальна даже тогда, когда она кажется интеллектуальной в *любом* внешнем отношении (то есть, в каждом отношении, проверенном тестом Тьюринга). Но, предлагая условие внутреннего [устройства], не предлагаете ли вы просто лингвистическое обусловливание слова «интеллект», его новое значение? *Обычно* способности ввода-вывода определяются как критерий интеллектуальности. Все, что вы предлагаете – это новую практику [использования слова «интеллект»], вводя *новый* критерий, включающий некоторые сведения о том, что происходит внутри, [а не только снаружи].

Ответ.

Джонс одновременно блестяще играет в шахматы против двух самых выдающихся в мире гроссмейстеров. Вы считаете его гением, пока не узнаете, что метод его игры следующий. Он ходит вторым в игре против мастера G1 и первым против G2. Он запоминает первый ход G1 против него, и потом делает точно такой же ход против G2. Далее, он ждет ответа G2 и делает такой же ход против G1, и так далее. Поскольку Джонс узнал об этом методе из комикса, игра Джонса не является признаком его интеллектуальности.

Как демонстрирует пример[10], особенность [традиционной] концепции интеллекта состоит в том, что система считается интеллектуальной в той степени, в какой её действия подражают поведению действительно ин-

10 Эти примеры были предложены Dick Boyd and Georges Rey в их комментариях к более ранним версиям этой статьи. Rey утверждал, что история с шахматами действительно происходила.

теллектуальной системы. При этом игра внешне интеллектуальной системы ошибочно принимается за признак её интеллектуальности. Поскольку действия моей машины суть *полностью* подражание, эти действия не могут служить основанием для приписывания машине интеллекта.

Дело в том, что хотя мы *обычно* приписываем интеллект системе путём анализа ее способностей по вводу-выводу [информации], наши обыденные представления по поводу этих способностей могут оказаться обманчивыми. Патнэм предложил рассматривать такого рода заблуждения в логике терминов естественных видов. В этой логике рассматриваются вопросы такого плана: даже когда [некий объект] X представляется стереотипным, на самом деле X может оказаться совсем не X-ом, так как его могут ошибочно рассматривать в сочетании с другими вещами, которые, в свою очередь, принадлежат к данному естественному виду и которые мы рассматриваем как X-ы. (Kripke, 1972; Putnam, 1975a) Если Патнэм прав, никто никого никогда не может обвинить в «изменении значения» термина естественного вида, в том числе и тогда, когда он утверждает, что нечто, удовлетворяющее стандартному критерию X-а, на самом деле X-м не является.

Возражение 3а.

Ваш ответ на последнее возражение кажется крайне подозрительным, особенно мнение о точке зрения Патнэма. Не является ли [явным] шовинизмом то положение, что система должна быть *похожа на нас* с научной точки зрения для того, чтобы быть интеллектуальной? Возможно, значение слова «интеллектуальность» будет принципиально иным для системы, способы обработки информации которой отличаются от наших способов. В равной степени верно и то, что наше понятие «интеллект» не попадает в сферу применения *ею* термина «шминтеллект». А кто сказал, что «интеллект» [интеллектуальнее], нежели чем «шминтеллект»?

Ответ.

Я не утверждаю, что здесь налицо *простой факт различия* информационных процессов в машине и в людях. Скорее моя точка зрения основана на *типе* существующих различий в информационных процессах. Моей машине не хватает типа «богатство» информационного процесса, необходимого для того, чтобы считаться интеллектуальной. Возможно, это «богатство» связано с абстрактными принципами решения проблем, обучения и т.д. Мне бы очень хотелось поговорить на тему о том, что такое это [интеллектуальное] «богатство». Но я выбрал менее честолюбивую задачу: предложить доступный пример, когда нечто *не обладает* этим «богатством», однако ведет себя так, как если бы на самом деле было бы интеллектуальным. Если кто-то предложит определение «жизни», неявным выводом из которого являлось бы признание живыми таких вещей, как скрепки для бумаги, вряд ли кто-то другой пытался бы даже его опровергнуть, указывая на абсурдность такого вывода, при всем при том, что у него

не нашлось бы подробного обоснования к понимаю того, что же такое «жизнь». Таким же образом кто-то может отвергнуть нео-ТТ и ему не нужно очень много говорить о том, что же на самом деле есть «интеллект», достаточно привести контрпример.

Возражение 4.

Вообразим, случилось следующее: люди, включая вас, стали обрабатывать информацию точно так же, как это делает одна из ваших машин. Будете ли Вы утверждать, что люди неразумны?

Ответ.

Я не очень уверен в том, что смогу что-то сказать об интеллекте человека, если меня убеждают в том, что информационный процесс внутри меня точно такой же, как у моей машины. В любом же случае, я не вижу какого-то *внятного и очевидно корректного* ответа на этот вопрос. Более того, ни один из правдоподобных ответов, которые я могу придумать, не совместим со всем тем, что я говорил до настоящего времени.

Возьмем для примера теорию референции, которая утверждает, что в силу связи между словом «интеллектуальность» и человеческим информационным процессом, последний является «интеллектуальным» *независимо* от собственной, внутренней природы этого процесса. Тогда если бы меня убедили, что человек обрабатывает информацию подобно моей машине, я вынужден был бы признать, что моя машина интеллектуальна. Но разве это не совместимо с моим утверждением, что моя машина *на самом деле* не интеллектуальна? Придираться ко мне словами: «А что было бы, если бы вы оказались тем-то» - это почти также, как если сказать атеисту: «А что было бы, если бы вы оказались Богом?» Атеисту пришлось бы согласиться, что если бы он был Богом, тогда бы Бог существовал. Но атеист может пойти на эту уступку, вовсе не отказываясь от атеизма. Если слово «интеллектуальность» надежно связано с человеческим информационным процессом, как это мне предлагают сделать, тогда моя позиция связана [не с голословным, а] с *эмпирическим утверждением*, что человеческий информационный процесс отличается от информационного процесса машины. И это утверждение поддерживается и здравым смыслом и эмпирическими исследованиями в когнитивной психологии.

Возражение 5.

Вы настаиваете на том, что мы обрабатываем информацию не так, как это осуществляет ваша машина. Что придаёт Вам уверенности в этом?

Ответ.

Я не понимаю, как кто-то может делать подобные замечания серьёзно, не в шутливой форме. У вас не возникнет затруднений при размышлении над моими доводами? Надо ли нам серьезно воспринимать идею, что кто-то когда-то давным-давно записал всё то, что я сейчас сказал и все ваши ответы на то, что я сказал, и затем всё это вставил в ваш [или мой]

мозг? Здравый смысл отшатывается от столь явной чепухи. Более того, возьмите любую публикацию любого журнала по когнитивной психологии. Вы найдёте описания попыток экспериментальных исследований механизмов нашего информационного процесса. Несмотря на грубость доказательств, они, в подавляющем большинстве, высказываются против идеи того, что наш информационный процесс реализует механизм построчного поиска.

Наши когнитивные процессы, без всякого сомнения, гораздо более механистичны, чем многие люди желают думать. Но есть огромная разница между тем, что наше существо более механистично, чем хотели бы думать люди, и тем, что наше существо есть машина описанного мною типа.

Возражение 6.

Комбинаторный взрыв делает вашу машину невозможной. Джордж Миллер давно показал (Miller et al., 1960), что существует порядка 10^{30} грамматических предложений длиною в 20 слов. Предположим (весьма произвольно), что из них 10^{15} семантически корректны. Тесту Тьюринга длительностью в час может понадобиться порядка 100 таких предложений. А это 10^{1500} строк - число, которое больше числа частиц во вселенной.

Ответ.

Мой довод требует только того, чтобы машина была *логически* возможна, а не практически, или даже номологически. Бихевиористские исследования, в основном, *абстрактные исследования*, и сложно помыслить, как концепции, подобные нео-ТТ, могут быть рассмотрены в ином свете. Можно ли считать э*мпирической гипотезой* то, что интеллект - суть способность выдавать осмысленные последовательности реакций, соответствующие входным последовательностям? Какого рода эмпирические подтверждения (отличные от лингвистических доказательств) можно привести в пользу такого утверждения? Если рассматривать нео-ТТ как концепцию интеллекта, тогда для опровержения этой концепции достаточно всего лишь *логической* возможности неинтеллектуальной системы, способной пройти тест Тьюринга.

Могут возразить: хотя концепция теста Тьюринга, несомненно, и не является *первостепенной* эмпирической гипотезой, но её можно считать квази-эмпирической. Против этого можно привести аргумент: отождествление интеллектуальности со способностью выдавать осмысленные ответы – это *второстепенный* принцип или второсортный закон эмпирической психологии. Подобного рода отождествление может быть предложено как рациональная реконструкция неопределенно общей концепции интеллекта, которая, возможно, найдёт полезное применение в будущей теории эмпирической психологии. В обоих случаях, пока что ни одно эмпирическое доказательство *не может* прямо поддержать нео-ТТ. Однако данную концепцию следует сохранить как часть перспективы, которая в будущем может быть целиком поддержана.

Такой ответ имел бы вес, если какой-нибудь сторонник концепции теста Тьюринга привёл бы любой, [пусть самый незначительный] довод в пользу того, чтобы считать концепцию ТТ плодотворной для создания эмпирических теорий, которые её хоть как-то учитывали бы. Почему мы должны в отсутствие такого довода (и, кроме того, при наличии примеров, которые предлагают антибихевиористская психология и подходы к искусственному интеллекту, отличные от тьюринговой концепии интеллекта), принимать нео-ТТ всерьёз как квази-эмпирическое утверждение?

Пока такого ответа достаточно. Добавлю, что моя машина, безусловно, может быть номологически возможной. Ничто в современной физике не запрещает возможности обнаружения в какой-нибудь части вселенной бесконечно делимой материи. Например (а для нашей части вселенной это происходит всякий раз), когда последняя «элементарная» частица оказывается в действительности неэлементарной и когда количество и разнообразие ее компонентов умножается (как произошло с кварками). Физики при этом склоняются к гипотезе, что *наша* материя не состоит из каких-то *действительно* элементарных частиц.

Предположим, что обнаружена та часть вселенной (где, возможно, обитаем и мы с вами), в которой материя бесконечно делима. В этой части вселенной невозможно установить верхнюю границу тому количеству информации, которая локализована в заданном ограниченном пространстве. В этой части вселенной моя машина может существовать. Её лёнты памяти [, на которые записывается бесконечно много строк,] небольшие по размеру, к примеру, с человеческую голову. Безусловно, можно вообразить, что там, где материя бесконечно делима, существуют творения самых разных размеров, включая существ размером с электрон. Эти существа могут, например, договориться с нами делать записи на ленту памяти, если мы будем истреблять их врагов в одном из наших ускорителей элементарных частиц.

Более того, даже если эта фантазия номологически невозможна, всё же остается неясным вопрос о её значимости в контексте возражения по поводу номологической невозможности нео-ТТ. Если нео-ТТ является эмпирически «второстепенной» закономерностью, то это – второстепенная закономерность *когнитивной психологии*, но не *физики*. Более того, ситуация может противоречить законам физики и не противоречить законам человеческой психологии. Например, в логически возможном мире гравитация может подчиняться закону обратно пропорционально кубу, вместо закона обратно-пропорционального квадрата и тогда законы *физики* будут различными. Однако законы *психологии* не обязательно будут нарушаться [в физически различных мирах].

Таким образом, если моя машина противоречит законам природы, то это – законы физики, но не психологии. То, как много информации может храниться в данном пространстве, и как быстро информация может передаваться из одного места в другое, зависит от таких физических факторов, как делимость материи и скорость света. Даже если только что описанные создания размером с электрон противоречат законам физики, они могут не противоречить законам психологии. То есть люди (которых не повредили психологические законы) могут [вполне благополучно] сосуществовать с маленькими созданиями[11].

Но если моя машина не противоречит законам человеческой психологии – если она существует в возможном мире, в котором законы человеческой психологии такие же, как здесь – тогда концепция нео-ТТ ошибочна в этом возможном мире, как ошибочна она и в нашем мире. Таким образом, нео-ТТ не может быть одним из законов человеческой психологии [ни первостепенным, ни второстепенным].

В итоге, нео-ТТ представляет либо разновидность абстрактной истины, либо разновидность психологического закона. И ошибочна эта концепция в обоих случаях.

Еще один завершающий момент: если осуществить различного рода модификации моей машины, то она окажется номологически возможной в более прямом смысле. Во-первых, можно сократить словарь теста Тьюринга до объёма упрощенного английского словаря, который включает всего лишь порядка 850 слов, а не сотни тысяч слов английского языка. Так же бытует мнение, что ограниченным английским языком можно обойтись для ведения нормальной беседы и выражения достаточно широкого круга мыслей. Во-вторых, способ вычисления основан на способе поиска по строкам. Версия машины поиска по деревьям, описанная ранее, позволяет избежать чрезвычайно большого количества повторений подстрок. Однако и эта версия ничуть не интеллектуальнее первой.

Более важным представляется то, что машина в той форме, какой я описывал, сконструирована слишком идеально – она работает *совершенно* [без сбоев]. Однако совершенное функционирование – это намного больше того, чего можно ожидать от любого человека, даже если не обращать внимания на инсульты, сердечные приступы и другие формы человеческих «сбоев». Люди со здравым рассудком могут неправильно понимать предложения, путаться. Любому же нормальному человеку, который участвует

11 Можно возразить, что поскольку информационный процесс гораздо более эффективен в мире, в котором материя бесконечно делима, нежели в нашем мире, законы мышления в том мире отличаются от законов мышления нашего мира. Но после этого возражения напрашивается вопрос. Если описанная мною машина поиска по строкам не может размышлять в *любом* мире (как я и утверждал), то номологическая разница, которая делает возможной данную машину - это разница в законах, которые влияют на *имитацию* мышления, а не на разницу в законах мышления.

в длительном тесте Тьюринга, вскоре должно наскучить это мероприятие, он станет невнимателен. Говоруны с самого начала будут нести бессмыслицу, иногда извиняясь, что не слушали собеседника. Многие люди на замечания собеседника будут отвечать свободными ассоциациями, нежели напрягать свой разум, давая вразумительные ответы. Некоторые будут постоянно жаловаться на непривлекательность этих бесконечных тестов Тьюринга.

Если создавать машину, которая функционировала бы так же хорошо, как действует *большинство* людей, проходящих тест Тьюринга, тогда, возможно, была бы разработана гибридная машина, в которой представлено относительно небольшое количество деревьев и которая использует набор хитрых приемов, подобных приёмам программы Вайзенбаума.

Можно было бы изобрести большое количество приемов сокращения загруженности памяти – но машина умнее от этого не станет. И неважно, насколько сложны приемы освобождения памяти, которые мы создадим - эти приемы дают отсрочку, но не предотвращают неотвратимость комбинаторного взрыва. И самым лучшим проектом такой машины станет замена памяти разумом. Так что для теста Тьюринга *неопределённой* длины гибридная машина окажется настолько большой, что попытка ее построить станет причиной провала [средств] в черную дыру.

Моя точка зрения такова: техническая изобретательность, сколь бы изощрённой она не мыслилась, выходит за рамки доступных приёмов, необходимых для моделирования речевых [или диалоговых] способностей человека. Такая изобретательность номологически невозможна.

Возражение 7.

Недостаток описанного Вами теста Тьюринга заключается в том, что тест представлен в форме экспериментального проекта, а не экспериментальной концепции. Ваш тест Тьюринга имеет *фиксированную длину*. Чтобы запрограммировать машину, программистам надо знать эту длину. В *адекватном* тесте Тьюринга длительность должна выбираться случайным образом для каждого сеанса тестирования. Короче, правомочность вашей критики нарушается тем, что вы помогаете людям.

Ответ.

Конечно, верно то, что способность моей машины пройти тест Тьюринга зависит от некоторой верхней границы длины теста. Но то же самое верно и для *людей*. Даже если мы даем человеку, скажем, двенадцать часов подумать между вопросом и ответом, ведь ему надо дать время поесть и поспать. Люди, помимо всего прочего, умирают. Лишь немногие люди смогут пройти тест Тьюринга, который длится девяносто лет, и ни один человек не сможет пройти тест Тьюринга длиной в пятьсот лет. Вы можете (если желаете) охарактеризовать интеллект способностью пройти тест Тьюринга случайной длины, но поскольку *люди не имеют такой возможности*, ваша характеристика не будет необходимым условием интеллекту-

альности. Даже если бы она была достаточным условием интеллектуальности (в чем я очень сомневаюсь – об этом см. ниже), достаточное условие интеллектуальности, которому *не удовлетворяют люди*, пользовалось бы малым спросом у сторонников нео-ТТ.

Даже если медицина когда-то и даст человеку бессмертие, люди всё-равно будут умирать, к примеру, из-за несчастных случаев. Например, имеется ненулевая вероятность, что даже в ходе нормального теплового обмена молекулы в двух частях одного тела станут двигаться в противоположных направлениях, разорвав, в конченом счёте, целое тело пополам. Конечно, вероятность избежать смерти от такого несчастного случая почти *всегда* равна нулю. Рассмотрим «период полураспада» людей в мире, в котором смерть откладывается так долго, как это физически возможно. (Период полураспада для людей, как и для атомов – это средний отрезок жизни, время, занимаемое, чтобы прошло полжизни). Машины моего типа могут быть запрограммированы, чтобы обеспечивать нео-ТТ длительностью в эти полжизни, и (в предположении, что они не более подвержены несчастным случаям, чем люди) их «половинный» отрезок функционирования будет длиться столько же, сколько он длится у среднего человека.

Возражение 7a.

Позвольте попробовать другой трюк. Когнитивные психологи и лингвисты утверждают, что когнитивные механизмы обладают «бесконечными способностями». Например, Хомский заявляет, что механизмам понимания языка предоставлена возможность понять предложения любой длины. Верхняя граница длины предложений, которые люди способны понять на практике образуется в результате помех - невнимания, скуки, ограниченности рассудка, конечности памяти и многих других феноменов, включая, конечно же, смерть. Практический аспект вуалируется идеализацией (т.е., игнорированием подобного рода «помехи») - в идеале мы всё же способны понимать предложения любой длины.

При подобного рода идеализации люди, вероятно, способны пройти тест Тьюринга любой длины. Однако ваша машина поиска по строкам, даже если применить эту идеализацию, таковой возможностью *не* обладает.

Ответ.

Вы думаете, что возразили моему утверждению. Вам это кажется. На самом деле вы *капитулировали* перед ним. Я готов признать идеализацию, в условиях которой мы, вероятно, и обладаем «бесконечной» способностью, отсутствующей у моей машины. Но для того, чтобы [применить и] описать идеализацию, вам придется потакать различного рода теориям когнитивных механизмов, а это для бихевиориста неприемлемо.

Рассмотрим возможную переформулировку нео-ТТ, в которой найдет место возражение по поводу идеализации. Переформулировка будет нечто чем-то вроде: «интеллектуальность = обладание механизмами лингвистической обработки, которые способны выдавать осмысленные ответы, снабжены бесконечной памятью, управляются мотивационными механизмами, устанавливают умеренно высокий уровень предпочтения для осмысленных ответов, «игнорируют» сигналы «останова» от центра эмоций и т.д.».

Заметьте, что для принятия такой теории следует провести различие между психическими составляющими и психическими механизмами. Я готов отступить и признать, что эти различия подкрепляются действительными эмпирическими предположениями. Однако мое отступление не пойдет на пользу противной стороне. Например, память может самым сложным образом переплетаться с лингвистическими механизмами и при таком представлении становится бессодержательным разговор о снабжении лингвистических механизмов неограниченной памятью. Вопрос ключевым образом решается так: для принятия «идеализированной» версии нео-ТТ придется призвать менталистские представления, а их ни один бихевиорист не примет.

Возражение 7b.

Я смогу Вам возразить и без использования менталистских представлений. Я буду идеализировать, отвлекаясь только от случаев неестественной смерти, т.е. смерти, которая произошла по причине катастрофы. На возражение 7 Вы ответили (и правильно), что если бы медицина дала человеку бессмертие, то всё равно, возможности «средней» машины, основанной на поиске по строкам, были бы сопоставимы с возможностями «среднего» [, но не бессмертного] человека. Однако заметьте, при условии исключения неестественной смерти, для каждого *конкретного* человека длительность прохождения теста Тьюринга становится безграничной. В отличие от этого, у любой конкретной строково-поисковой машина есть верхний предел возможностей.

Ответ.

Определяющим в вопросе, как долго может продолжаться тест Тьюринга, является не то, как долго мы живем, а природа наших когнитивных механизмов и [способы] взаимодействия их с другими психическими механизмами. Допустим, у нас нет механизмов «стирания» информации, хранимой в долгосрочной памяти. (Так ли это – неизвестно.) Если мы не сможем «стирать», тогда настанет момент, когда наша память – а она конечна – «израсходуется» и способность нормально общаться иссякнет.

Если бихевиорист идентифицирует интеллект со способностью неограниченно долго продолжать тест Тьюринга, отвлекаясь при идеализации от неестественной смерти, тогда в его представлении люди также могут оказаться неразумными. Более того, даже если окажется, что люди удовлетворяют такому условию, оно не может рассматриваться в роли *необходимого* условия интеллектуальности. Существа, стареющие на протяжении двух столетий, так как у них отсутствуют «стирающие» механизмы, могут, тем не менее, до старости быть разумными.

Конечно, бихевиорист может избежать этой трудности посредством дальнейшей идеализации, явно упоминая механизмы стирания в своем определении интеллектуальности. Но это вновь его потянет в трясину менталистского болота, образовавшегося в предыдущем ответе.

Стоит добавить, что даже если у нас и есть механизмы «стирания» и даже если исключить все случаи неестественной смерти, то мы всё равно обладаем *конечной* памятью.

Мой строковый поисковик, работающий при условии конечной памяти, может действовать также, как действует человек, проходя неограниченно длительный тест Тьюринга. Предположим к примеру, что человеческая память не способна вместить в себя вопросы-ответы из диалога, длящегося более двухсот лет. Тогда можно создать версию машины, осуществляющей построковый поиск, превратив оригинальную машину в *циклический поисковик*. Эта версия может функционировать неограниченное время. Вместо «линейных» строк диалога он содержит циклические строки, окончания которых состоят из начал после, скажем, тысячи лет диалога. Конструкция таких циклов потребует гораздо больше изобретательности, чем конструирование обычных строк. Даже если это и возможно, способ функционирования этой машины будет казаться постоянно повторяющимся тому существу, объем памяти которого намного превосходят объем памяти нашей машины и человеческий диалог такому существу скоро наскучит.

И ещё один, завершающий ответ на возражение о «безграничном тесте Тьюринга». Рассмотрим версию моей машины, в которой программисты просто добавляют и добавляют строки. Когда им нужна новая лента, они используют уже пройденную ленту[12]. Заметьте, *логически* возможно [автоматическое], без участия программистов, формирование строк, которые удлиняются сами по себе. Таким образом, даже неспособность неограниченно долго продолжать тест Тьюринга, является *логически* достаточной для интеллектуальности.

Продолжая эту тему, рассмотрим бесконечно делимую материю, упомянутую в ответе на возражение 6. Для *человеко-размерной* машины, осуществляющей поиск по строкам, логически возможно, а может, даже номологически возможно, включать в свой состав существ постоянно уменьшающегося размера, которые без конца работают и работают, производя все более длинные и длинные записи на ленты.

Конечно, ни одна из двух только что упомянутых машин не имеет *конкретной* программы, [применяемой для жёстко фиксированных условий диалога]. И так как программисты никогда не воспринимают стимул

12 Эта машина достигла бы большего, если бы программисты имели возможность отказываться от строк, которые становятся бесполезными в зависимости от направления беседы. (В версии машины поиска по деревьям это эквивалентно методу отсечению ветвей)

непосредственно, то отвечает *машина*, а не программисты. Сравните эти машины с печально известными «машинами» прошлых столетий, в которых незаметно прятался человек, слушал вопросы и отвечал на них [посредством механизмов].

Возражение 8.

Ранее Вы отмечали, что нео-ТТ широко распространена среди исследователей искусственного интеллекта. Однако Ваша машина не может опровергать все точки зрения на ИИ (искусственный интеллект). Так, например, Ньюэлл и Саймон, утверждают: «задача интеллекта… предотвращать угрозу экспоненциального взрыва при поиске [решения задачи]» (Newell and Simon, 1979) (При экспоненциальном взрыве при поиске решения задачи добавление одного шага решения требует увеличения вычислительных ресурсов в 10 раз, добавление двух шагов требует увеличения вычислительных ресурсов в 10^2=100 раз, добавление трех шагов требует увеличения вычислительных ресурсов в 10^3=1000 раз и т.д.). Поэтому сторонники искусственного интеллекта вполне обоснованно смогут представить свою собственную версию нео-ТТ следующим образом:

> Интеллект – это способность выдавать осмысленные последовательности ответов на стимулы до тех пор, пока не произойдёт экспоненциальный взрыв при поиске [решения][13].

Ответ.

Сначала обратим внимание, что для сторонника оригинальной [концепции] нео-ТТ придерживаться измененной версии – это значит капитулировать перед психологизмом, который я защищаю. *Исправленная* нео-ТТ пытается избежать проблемы, которую я поставил, включив в [определение интеллекта] *условие [рассмотрения] внутреннего процесса* (хотя бы, упоминание об этом процессе). Исправленная версия упоминает о нём – что он не будет взрывоопасным. Таким образом, измененная нео-ТТ частично характеризует интеллект в плане её внутренних этиологических показателей. Поэтому исправленная нео-ТТ – это психологистская концепция.

Поскольку исправленная нео-ТТ улучшает первоначальную концепцию нео-ТТ исключительно в этом одном отношении (вовлекает в определение внутренние процессы), ей присущи многочисленные недостатки. Один проистекает из-за двусмысленности фраз, таких как: «предотвращает экспоненциальный взрыв поиска», что может быть понято в смысле: «*избегает* экспоненциальных взрывов в целом» (использует методы, в которых не задействованы вычислительные ресурсы, экспоненциально увеличи-

13 Я признателен Даниэлу Деннету за столь сильное возражение. Он предложил это возражение, когда выступал в качестве оппонента к более ранней версии этой статьи в Университете Цинциннати на Конференции по философии психологии в 1978 г. Деннет заявляет, что готов защищать нео-ТТ в её исправленной версии.

вающиеся с ростом «длины» задачи), или, например, в смысле: «*отсрочивает* экспоненциальный взрыв достаточно долго» (использует методы, требующие вычислительных ресурсов, которые экспоненциально возрастают с увеличением «длины» задачи, «длина» задачи при этом достаточно коротка, что определяет доступность фактически требуемых ресурсов).

Если имеется в виду «отсрочка», то мой контрпример прекрасно справляется с новой формулировкой нео-ТТ, потому что, как я указывал ранее, моя машина или её версия способны отсрочивать комбинаторный взрыв достаточно долго для прохождения тест Тьюринга приемлемой длины.

С другой стороны, если имеется в виду *избежание* всех комбинаторных взрывов, тогда изменённая нео-ТТ может характеризовать людей как неразумных [существ]. Хотя возможно, что человеческие механизмы, реализующие информационные процессы, как и механизмы многих систем ИИ, преуспели не потому, что избегают всех комбинаторных взрывов, а лишь потому, что для практических целей *отсрочивают* комбинаторные взрывы достаточно длительное время.

В итоге, изменённая нео-ТТ сталкивается с дилеммой. Если имеется в виду отсрочивание комбинаторного взрыва, тогда мою машину можно считать интеллектуальной. Если имеется в виду избежание всех возможных комбинаторных взрывов, тогда *нас* (и других разумных организмов) нельзя считать разумными.

Более того, предложенное изменение нео-ТТ – это чистое ad-hoc добавление. Проблема с подобного рода ad-hoc–контрпримером состоит в следующем: нет никакой уверенности, что кто-то иной не предложит другой контрпример, который, в свою очередь, потребует иного ad-hoc–маневра.

Вернёмся к началу [ответа на возражение], прочертив контуры устройств, которые обладают осмысленными зависимостями ввода-вывода, но которые, однако, неразумны.

Представим компьютер, который моделирует ваши ответы на вопросы-стимулы, вычисляя *траектории* всех элементарных частиц вашего тела. Такая машина стартует с подробного описания положений, скоростей и состояний всех элементарных частиц (для удобства можно использовать механику Ньютона). Она вычисляет изменения состояния вашего тела как функцию от этих начальных условий и сил, задействованных вашими сенсорными механизмами. Примем как безусловное то, что для реализации теста Тьюринга особенно важен эффект освещения монитором ваших пальцев на клавиатуре. Теперь, хотя это и потребует прояснения, я выскажу мнение, что машина, которая вычисляет траектории ваших элементарных частиц, однако, не является интеллектуальной. Да, она способна управлять роботом, который может вести себя в точности так же, как ведёте себя вы, причём в любой ситуации. Робот может вести себя так, как,

например, ведёте себя вы, когда философствуете. Однако на самом деле этот робот не занимается философией. То, что *он* делает – это вычисляет траектории элементарных частиц, чтобы подражать вам, занимающемуся философией.

Возможно, то, что я здесь описал, номологически невозможно. Даже если бы Бог задал положения и скорости всех частиц нашего тела, то всё равно ни один компьютер не смог бы сосчитать столь сложные взаимодействия, несмотря на то, что пользуется [простыми соотношениями] механики Ньютона.

Тем не менее, заметим одну особенность, благодаря которой эта машина может быть лучше машины, описанной в данной статье. А именно: если она может имитировать что-то в течение часа, она может оказаться способной имитировать это в течение года или десятилетия на базе той же самой аппаратуры. Для продолжения процесса моделирования просто требуется снова и снова решать одни и те же уравнения. Ввиду широкого ряда разновидностей типовых решений, постоянное решение уравнений не приведёт к экспоненциальному взрыву при поиске. А если нет экспоненциального взрыва, то исчезает и необходимость условия, добавленного в данном возражении при определении изменённой нео-ТТ. Мы остались лишь с проблемами номологической возможности из Возражения 6.

Идея только что обрисованной машины может быть воплощена в другой машине, которая ближе к номологической возможности - в машине, которая моделирует вашу *нейрофизиологию*, а не физику ваших элементарных частиц. В этой машине будут репрезентированы: некоторая адекватная нейрофизиологическая теория (скажем, эта теория появится в некотором отдалённом будущем) и спецификация текущих состояний ваших нейронов. Машина моделирует вас путем вычисления изменений в состояниях ваших нейронов. Если пойти далее, то все же лучшей с точки зрения номологической возможности будет машина, моделирующая вашу *психологию*. В ней репрезентируется некоторая адекватная психологическая теория (отдаленного будущего) и спецификация текущих состояний ваших психологических механизмов. Моделирование вас осуществляется путём вычисления изменений в состояниях этих механизмов при учёте исходных состояний на сенсорных входах. И вновь, даже если нет экспоненциального взрыва, такая модификация машины ничего не дает.

Я говорил, что эти три устройства *сомнительно* неинтеллектуальны, но, поскольку я не располагаю временем для детального доказательства, придется оставить незавершенной данную часть вопроса. Я лишь кратко очерчу контуры аргументации.

Рассмотрим устройство, которое моделирует вас, используя теорию ваших психологических процессов. Это – робот. Он выглядит и ведет себя так, как ведёте себя вы в любой ситуации стимуляции. Вместо мозга у робота – компьютер. В компьютере – описания способов функционирования

ваших когнитивных механизмов. На вход к вам поступает определенная информация, вы размышляете над ней и выдаете определенную информацию на выходе. Если ваш робот-двойник [doppelganger] получает на входе такую же информацию, какую получаете и вы, его приемник преобразует эту информацию в дескрипции (описания) входа. Компьютер использует дескрипции ваших когнитивных механизмов, чтобы проследить траекторию ваших размышлений. Далее он пересылает дескрипции ваших ответов механизму, который обуславливает действия робота. Не совсем очевидно, что процесс манипуляции роботом дескрипциями ваших размышлений сам по себе является *размышлением*. Еще менее очевидно, что манипуляции дескрипциями вашего опыта и ваших эмоций у робота сами по себе являются опытом и эмоциями.

Чтобы слегка напрячь вашу интуицию в решении этого вопроса, заменим компьютер, манипулирующий дескрипциями в голове робота - вашего двойника, на очень маленького *разумного* человека. Этот человечек говорит только по-китайски. У него есть руководство (на китайском), описывающее ваши психологические механизмы. Вы получаете входную информацию: «Кто ваш любимый философ?» Немного подумав, отвечаете: «Гераклит». Ваш робот-двойник, с другой стороны, содержит механизм, который преобразует задаваемый вопрос в дескрипцию звука. Затем маленький человечек делает вывод, что вы в ответ на этот вопрос издаёте шум: «Гераклит». И заставляет «гортань» робота издать этот шум. Робот мог бы моделировать зависимости ваших входов-выходов (и, до известной степени, моделировать ваши внутренние информационные процессы), даже не смотря на то, что человечек внутри робота не понимает ни вопроса ни ответа. Возникает впечатление, что робот, моделируя вас, думает вашими мыслями. Однако сам по себе он этими мыслями не думает. Возвращаясь к случаю, когда у робота внутри – компьютер, который вместо человечка манипулирует дескрипциями, зададимся вопросом - почему мы должны полагать, что робот обладает [ментальными состояниями] или что в нём протекают какие-то мыслительные операции?

Анализ машины построчного поиска, на основе которой строилась данная статья, показывает, что поведение интеллектуально только тогда, когда оно *не является* продуктом информационного процесса описанного нами типа. Возвращаясь к примеру с марсианами, приведённому в начале статьи, я предостерегаю от вывода, что можно каким-то образом положительно определить тип того информационного процесса, который лежит в основе всего разумного поведения (помимо того, что этот процесс должен обладать хотя бы минимальной степенью «богатства»). Однако всё то, что было сказано в связи с марсианами и с машиной построчного поиска, оставило открытым следующий вопрос. Хотя и не существует некоторого единого естественного информационного процесса, лежащего в основе всего интеллектуального поведения, но, может быть, существует некий информационный процесс, общий для всех *не*интеллектуальных существ, справляющихся с тестом Тьюринга (а именно, существует достаточно простой процесс, оперирующий памятью большого объёма)? То есть такая машина предполагает существование информационного процесса, единого для всех неразумных систем. Но эта возможность сомнительна.

Литература

Block, Ned. 1978b. 1978a. Reductionism. In *Encyclopedia of bioethics*. New York, NY: Macmillan.

Block, Ned. Troubles with functionalism. In *Perception and cognition: Issues in the foundations of psychology*, ed. C. W. Savage, vol. 9 of *Minnesota Studies in the Philosophy of Science*. Minneapolis, MN: University of Minnesota Press.

Block, Ned.1980. Are absent qualia impossible? *Philosophical Review* LXXXIX.

Boden, M. 1977. *Artificial intelligence*. New York: Basic Books. Chisholm, Roderick.1957.

Perceiving, chap. 11. Ithaca, NY: Cornell University Press.

Dennett, Daniel. 1978. *Brainstorms*. Cambridge, MA: MIT Press. Descartes, René. 1985. *The philosophical writings of René Descartes*. Cambridge, England: Cambridge University Press. Translated by John Cottingham, Robert Stoothoff, and Dugald Murdoch.

Donne, John. 1980.*Paradoxes and problems*. Oxford, England: Clarendon Press. Edited with an introduction and commentary by Helen Peters.

Dummett, Michael. 1976. What is a theory of meaning (II). In *Truth and meaning*, ed. G. Evans and J. McDowell. London: Oxford University Press.

Fodor, Jerry. 1968. *Psychological explanation*. New York: Random House. Gunderson, Keith. 1964.

The imitation game. *Mind* 73(290): 234-245.

Kripke, Saul. 1972. Naming and neccesity. In *Semantics and natural language*. Dordrecht, Holland: Reidel.

Lycan, William. 1979. New lilliputian argument against machine functionalism. *Philosophical Studies* 35.

Block, Ned. 1981. Form, function, and feel. *Journal of Philosophy* LXXVIII: 24-50.

Miller, George, Eugene Galanter, and Karl H. Pribram. 1960. *Plans and the structure of behavior*.

New York, NY: Holt, Rinehart, and Winston.

Newell, Alan, and Herbert Simon. 1979. Computer science as empirical inquiry: Symbols and search. *Communications of the Association for Computing Machinery* 19.

Putnam, Hilary. 1975a. The meaning of `meaning'. In *Language, mind, and knowledge*, ed. Keith Gunderson, vol. 7 of *Minnesota*
Studies in the Philosophy of Science. Minneapolis, MN: University of Minnesota Press.
Block, Ned. 1975b. *Mind, language, and reality*. Cambridge, England: Cambridge University Press.
Rorty, Amélie. 1979. *Philosophy and the mirror of nature*. Princeton, NJ: Princeton University Press.
Ryle, Gilbert. 1949. *The concept of mind*. London, England: Hutchinson.
Schank, Roger C., and Robert P. Abelson. 1977. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*.
Hillsdale, NJ: Lawrence Erlbaum Press.
Searle, John R. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3: 417-457.
Shoemaker, Sydney. 1975. Functionalism and qualia. *Philosophical Studies* 27.
Block, Ned. 1982. The missing absent qualia argument--a reply to block. Check. Block has s forthcoming.
Turing, Alan M. 1950. Computing machinery and intelligence. *Mind* LIX(236): 433-460.
Weizenbaum, Joseph. 1966. Eliza -- a computer program for the study of natural language communication between man and machine. *Communication of the Association for Computing Machinery* 9(1): 36-45.
Block, Ned. 1976. *Computer power and human reason*. San Francisco, CA: W. H. Freeman and Co.

Селмер Брингсйорд, Пол Беллоу, Дэвид Феруччи

Творчество, тест Тьюринга и улучшенный тест Лайвлейс

Перевод Алексея Ласточкина

Перевод выполнен по работе: Selmer Bringsjord, Paul Bello, David Ferrucci, 2000, http://www.rpi.edu/~faheyj2/SB/SELPAP/DARTMOUTH/lt3.pdf

Основные положения

Многие утверждают, что Тест Тьюринга (ТТ) – это один из путей проверки наличия у компьютера такого таинственного феномена как мысль и сознание. К сожалению, попытки создания вычислительных систем, способных пройти ТТ (или, по крайней мере, его менее строгие версии), выродились в узкоспециализированную сферу манипуляции символами, имеющую цель обмануть судью тем или иным способом. Разработчики таких систем отлично понимают, что всего лишь пытаются надуть людей, которые, взаимодействуя с машиной, верят, что у нее действительно есть разум. Вся суть проблемы заключается в том, что сами особенности ТТ способствуют появлению таких ловкачей. Нам кажется, что правильный подход заключается в задании определенных эпистемических отношений между искусственным агентом **A**, продуцируемым им выходом ***o*** и человеком **H**, который создаёт искусственного агента**.** Такие отношения, грубо говоря, возникают в тех случаях, когда **H** не может вычислить, как **A** вывел ***o***. ТТ при условии задания этих ограничений будем называть тестом Лавлейс, в честь леди Лавлейс, которая считала, что только при условии *самостоятельного создания* компьютерами *уникальных, неповторимых вещей* их следует считать обладающими разумом. Однако компьютеры на это не способны.

1. Введение

Вероятно, вы знаете о предсказаниях Тьюринга – он утверждал в своей работе «Вычислительные машины и интеллект (1950)», что к концу столетия компьютеры станут настолько умными, что, общаясь с ними на расстоянии (сегодня это означает, к примеру, с использованием средств электронной почты), мы не сможем отличить их от людей. Такие компью-

теры смогут проходить то, что сейчас называется Тестом Тьюринга[1]. Однако 31 декабря 1999 года пришло и ушло, праздничная пиротехника отгремела, а факт остается фактом: ИИ так и не смог создать компьютер с разговорными способностями ребенка, только начавшего ходить.

Неприятное, конечно, заключается не в этом, а в том, что движение к мечте Тьюринга идет только за счет жульничества, пусть талантливого и умелого, но жульничества. Например, создатели искусственных агентов, которые сейчас соревнуются в современных версиях ТТ, отлично понимают, что всего лишь пытаются обмануть людей, взаимодействующих с системой, заставляя их поверить в то, что она действительно обладает разумом. Искусственный агент в этой ситуации полностью подпадает под общеизвестный Аргумент китайской комнаты (Chinese Room Argument) Серля (АКК): эти системы таким образом *разработаны* своими создателями, что когда они совершенно бессмысленно начинают манипулировать символами, «наивные» наблюдатели воспринимают результаты их работы как указание на стоящий за ними разум, но за этим стоит только сборник инструкций и правил. На самом деле здесь играют люди-разработчики против людей-судей. Промежуточные же вычисления в ходе тестирования во многих отношениях являются крайне примитивными.

Нам кажется, что более правильный подход заключается в наложении определенных эпистемических отношений между искусственным агентом **A**, выводимыми им данными ***o*** и человеком-разработчиком **H** этого искусственного агента. Эти отношения, грубо говоря, характерны для случаев, когда **H** не может вычислить, как **A** вывел ***o***. Такой Тест Тьюринга следует назвать тестом Лавлейс, в честь леди Лавлейс, которая была убеждена, что только при условии *самостоятельного создания* компьютерами *неповторимых вещей* их можно считать наделёнными разумом.

Ниже даётся краткое описание статьи. В разделе 2 более детально рассматривается возражение Лавлейс и два ответа: первый ответ Тьюринга (1964) и робототехника Ганца Моравеца (Hans Moravec), который придерживается тьюринговых позиций. Как вы увидите, оба ответа мало убедительны. Также в этом разделе опровергается второй ответ Тьюринга на возражение Лавлейс, в котором он утверждает, что компьютеры его удивляли. Третий раздел посвящен изучению параметров Теста Лавлейс. В четвёртом разделе мы подвергаем тесту Лавлейс трёх искусственных агентов. Данное трио принадлежит к категории «творческих» программ. (Ясно, что компьютер способный «творить», обладает большими шансами обойти возражение Лавлейс, нежели любая система, участвующая в ТТ и слепо выполняющая простые манипуляции с символами). Первый из трех рассмотренных агентов называется BRUTUS. Его разработали Брингсйорд и Ферруччи (Bringsjord and Ferrucci). Две другие системы: LETTER SPIRIT и COPYCAT разработаны Дугласом Хофштадтером (Douglas Hofstadter), который, возможно, является ведущим в мире специалистом по вопросу

компьютерного творчества. В пятой части доказывается несостоятельность третьего ответа Тьюринга (по Turing 1964) на возражение леди Лавлейс. Тьюринг обращается к собранным на основе нейронных сетей машинам–детям («child machines»). В шестой части мы обращаемся, на основе всего выше перечисленного, к самой возможности создания систем, которые смогут «прорвать» границы простой манипуляции символами и пройти ТЛ – это машины-оракулы (oracle machines). Так же мы показываем, что они, тем не менее, подпадут под АКК Серля: они все равно не будут способны ни на что большее, помимо бездумного манипулирования символами в соответствии с программами и, следовательно, не выполнят основные требования ТЛ, а именно: самостоятельно мыслить. В последней, седьмой части мы коротко описываем мораль сей басни: для того чтобы пройти ТЛ и самостоятельно мыслить, у системы должна иметься возможность обрести «свободную волю» в соответствии с собственной «мотивацией» (т.е. иметь место “агентная причинность»). Если это так, то создание машины, способной пройти ТЛ, будет весьма сложной задачей.

2. Возражение Лавлейс с самого начала

Возражение Леди Лавлейс, возможно, было наиболее сильным из всех возражений, выдвинутых против ТТ. Перефразированное, возражение может звучать примерно так:

> Компьютер не может самостоятельно творить, т.к. творчество требует, как минимум, *изобретения* чего-либо *нового*. Но компьютеры не изобретают ничего нового; они всего лишь делают то, что мы приказываем им делать посредством программ.[2]

Что на это отвечает Тьюринг? Здесь мнения расходятся: в лучшем случае ответ его считают загадочным, в худшем – некомпетентным. Леди Лавлейс ссылается (в своих мемуарах) на аналитическую машину Ч. Бэббиджа (Babbage’s Analytical Engine). У этой машины Тьюринг с готовностью признает отсутствие возможности, как он выразился, «самостоятельно мыслить».

Таким образом, он признает, что так как возражение Лавлейс относится к машине Бэббиджа, то возражение правомочно. Но такие способности, как самостоятельное мышление или изобретение нового, как полагал Тьюринг, должны появиться у *других* специализированных машин. А такие машины (обозначим её через *M*), т.е. у компьютеры, появились уже *после смерти* Лавлейс. Однако Тьюринг указывает, что аналитическая машина может быть представлена *универсальным* цифровым компьютером, так что если её правильно запрограммировать, то она будет способна моделировать любую машину, в том числе, машину *M*. Но такая имитация придаёт аналитической машине способность к изобретению.

Данная аргументация просто не проходит по той простой причине, что вряд ли Лавлейс могла бы принять само допущение существования машины *M*. О какой машине утверждал Тьюринг? Какая машина подходит под список требований Лавлейс? Он так и не обнародовал этих требований, и факт остается фактом: лучшее, что он и его последователи смогли предложить – это машины, наивысшее достижение которых заключается в решении только арифметических задач.

После обсуждения аналитической машины, Тьюринг неожиданно трактует возражение Лавлейс, как утверждение, что компьютеры всегонавсего «не удивляют» (стр. 21-22 Turing 1964). Далее он ссылается на то, что считает непосредственным опровержением, а именно, на своё суждение «машины весьма часто меня удивляли».

Ответ Тьюринга не так давно был переработан Гансом Моравецем (Hans Moravec), который убеждён в том, что к 2040 г. будет преодолен не только ТТ, но и тотальный ТТ (ТТТ). (Согласно Стивену Харнаду (Stevan Harnad, 1991) робот проходит ТТТ, если он ничем – ни внешним видом, ни поведением, ни речью не отличаться от человека). Вот что Моравец заявляет:

> Леди Лавлейс, одна из первых программистов, никогда не имела в своем распоряжении работоспособного компьютера, который мог бы показать последовательность выполнения её программ. Современные программисты знают об этом вопросе много больше. Почти каждая новая программа ведет себя просто ужасно до тех пор, пока не будет потрачено много сил и времени на отладку и все равно ее так никогда и не удается приручить полностью. В то же время информационные системы, такие как системы с режимом разделения времени (time-sharing systems) или сети отличаются еще более диким поведением, причиной которого являются непредвиденные взаимодействия компьютеров, сливание потоков информации и попытки получения доступа.

Это крайне слабое возражение. Конечно, мы все знаем, что компьютеры иногда осуществляют непредусмотренные программистами действия. Однако это происходит лишь в силу того, что мы недостаточно старательны, дисциплинированы и предусмотрительны. Если мы и говорим о сбоях в оборудовании, то это из-за того, что некоторые микропроцессы протекают непредвиденным образом. Непредсказуемость процесса в этом случае никак *не* следует из того, что вычислительная система хочет самостоятельно *создать* что-то *оригинальное*. Чтобы разобраться, в чем дело, рассмотрим сборку Тойоты. Предположим, монтируя бампер, робот случайно на его место устанавливает запасное колесо вместо того, чтобы вложить его в нишу багажника, как предусмотрено инструкцией. Положим, что причиной этого явился «удачный» низкоуровневый сбой оборудования

или ошибка, возникшая по небрежности программиста. Далее допустим цепь счастливых случайностей: новое положение запаски становится для дизайнеров первым победным шагом к созданию наполовину традиционного седана, наполовину удобного спортивного автомобиля. И что же, честь создания нового автомобиля мы отдадим сломанному роботу? Конечно, нет.

Схожую ситуацию мы имеем, когда рассматриваем специфические взаимодействия между программистами и плохо разработанными программами, которые имели в виду Тьюринг и Моравец. Так как наш коллектив авторов постоянно программирует и, более того, мы профессионально обучены этому, нам не составит особого труда оценить данную ситуацию.

На наш взгляд, сюрпризы по вине простых синтаксических ошибок это не то же самое, что Тьюринг и Моравец имели в виду. Чтобы это увидеть, положим, что в одной большой программе *P* мы делаем попытку встроить простую функцию утроения заданного натурального числа на языке ЛИСП, создав такой код:

```
(defun triple(n)
(* m 3))
```

Теперь предположим, что в строке вызова («Lisp prompt >«) мы наберем (triple 6) и вдруг неожиданно получаем 75 (конечно, на самом деле такая обычная функция вызывается через другую функцию, но, чтобы упростить ее описание, допустим, что обращение идет напрямую). Очевидно, ceteris paribus[1] это нас удивит. В чем дело? Несмотря на то, что параметр функции triple задан как *n* в списке аргументов в дефиниции функции, в теле функции это *m*, т.е. *m*, а не *n* умножается на 3. При этом мы не помним, что ранее *m* объявили как глобальную переменную и присвоили ей значение 25. Отсюда и ошибка.

Сюрприз подобного рода, по всей видимости, не совсем то, что Тьюринг и Моравец имели в виду. Они, возможно, имели в виду некую разновидность *семантической* ошибки. Как можно определить такую ошибку? Рассмотрим пример, который станет основой обсуждения системы BRUTUS (пример использован для формирования множества разных точек входа (Bringsjord & Ferrucci 2000)).

Предположим, некто Билл пытается создать систему, способную обосновать идеи, которыми руководствовался какой-то человек, предавая другого человека. И, положим, используя весьма наивную логику первого порядка, Билл программирует «предательство», т.е. задаёт (не совсем уж и невероятное) определение:

---

[1] При прочих равных условиях (лат.)

Def_B 1 Агент s_r предаст агента s_d при таком положении дел *p*, что:

1. s_d хочет, чтобы *p* произошло;
2. s_r верит, что s_d хочет, чтобы *p* произошло;
3. s_r согласен с s_d что *p* должно произойти;
4. s_r добивается чтобы *p* не произошло;
5. s_r верит, что s_d верит в то, что s_r добивается, чтобы *p* произошло.

Этот код можно записать (при тех же условиях) на примере Дейва и Селмера (Dave and Selmer) при описании ситуации, когда Селмер никогда не сможет предать Дейва.

```
set(auto).
formula_list(usable).
% DefB-2 in {\sc otter}:

all x y (Betrays(x,y) <->
(exists z (Wants(y,z)
 Believes(x,Wants(y,z))
 Agrees(x,y,z) &
 IntendsNot(x,z) &
 Believes(x,Believes(y,Intends(x,z)))))).
% Pretend facts of the case:
Wants(adave,agraduate).
Believes(aselmer,Wants(adave,agraduate)).
Agrees(aselmer,adave,agraduate).
IntendsNot(aselmer,agraduate).
Believes(aselmer,Believes(adave,Intends(aselmer,agraduate))).
% Assumption for indirect proof:
-Betrays(aselmer,adave).
end_of_list.
```

Это – действующий код программы автоматического доказательства теорем, известной как OTTER.[2]

Заметьте, что этот код таким образом использует логику первого уровня, что никогда не будет работать. Чтобы понять почему, обратимся к формуле:

$$\forall x(P(x) \rightarrow Q(P(x))).$$

Эта формула бессмысленна по стандартам грамматики логики первого порядка. Антецедент *P(x)* в формуле под квантором всеобщности должнен принимать либо истинное либо ложное значение. Если он истинен (при подстановке какого-то конкретного *x*), тогда консеквент справа от → должен быть истинным. Положим, что *P(x)* некая структурная формула, а не выражение. Тогда консеквентом будет структурная формула, аргумент которой сам является структурной формулой. А это грамматически неверно и бессмысленно в логике первого порядка.

[2] OTTER может быть найден на http://www-unix.mcs.anl.gov/AR/otter/

Но, тем не менее, давайте предположим, что Билл как-то заставил систему исполнять код. Он ожидает сообщения об ошибке. Но что же произойдёт на самом деле? Противоречие будет найдено, и Билл очень удивится, найдя, что его система с большой вероятностью доказала, что Селмер предаст Дейва. В этом и проявляется семантическая ошибка. Собственно, программный «глюк» происходит оттого, что OTTER переинтерпретировал часть кода таким образом, что believes является одновременно и предикатом и функцией.

Чтобы разобраться в том, как это происходит, рассмотрим случай подачи на вход простой формулы вместе с условием *P(a)* (где *a* константа) и предполагаемого косвенного доказательства $\neg Q(P(a))$. Файл на входе:

```
set(auto).
formula_list(usable).
all x (P(x) -> Q(P(x))).
P(a).
% Assumption for contradiction:
-Q(P(a)).
end_of_list.
```

А вот доказательство на выходе:

```
---------- PROOF ----------
1 [] -P(x)|Q(P(x)).
2 [] -Q(P(a)).
3 [] P(a).
4 [hyper,3,1] Q(P(a)).
5 [binary,4.1,2.1] $F.
------------ end of proof -------------
```

Мы видим всё ту же семантическую ошибку. OTTER так и не показал, что же в данном случае искал наивный программист. P – функциональный логический элемент в строке 4 доказательства и, в то же время, утверждение в строке 3. Тот же феномен проявляется и в случае с Биллом.

Теперь, когда у нас имеется образец семантической ошибки, которая дает столько поводов «удивляться», зададимся вопросом: Будет ли основанием для утверждения о том, что система способна создать нечто оригинальное, факт удивления Билла машиной? Ясно, ни коим образом.

Как Тьюринг и Моравец могли допустить такую оплошность? Если размышлять доброжелательно и снисходительно к ним, насколько это возможно, то этот вопрос приводит нас к догадке, что «чувство удивления», которое они имели в виду, должно иметь более глубокие основания, нежели чем их объяснения.

Их удивление можно сравнить с удивлением прохожего, когда кто-то неожиданно вдруг выпрыгивает из-за угла дома и громко закричит при этом. Прохожий может удивиться до глубины души, хотя, в принципе, ваше поведение не такое уж экстравагантно. Для раскрытия подобного рода формы «удивления», которое испытывал Тьюринг, когда его удивляли программы, предложим версию ТТ. Назовём эту версию тестом Лавлейс (ТЛ).

3 Тест Лавлейс

Перед началом рассмотрения особенностей работы ТЛ, начнем со сценария, который близок к работе Брингсйорда и Ферруччи (Bringsjord and Ferrucci), которые много времени посвятили созданию агентов, генерирующих рассказы. Предположим, что Джонс – человек и сторонник ИИ, пытается создать искусственного компьютерного агента *A*, который способен рассказать что-то новое, по сути, сотворить (именно в понимании Лавлейс – то есть создать на самом деле оригинальный новый рассказ). Допустим, что Джонс запускает *A*, и тот выдаёт потрясающее художественное произведение *o*. Мы утверждаем, что, если наше главное условие выполнено, т.е. Джонс не может объяснить, как *A* сгенерировал *o*, и нет оснований считать, что успех достигнут по счастливому стечению обстоятельств, в ходе сбоя и т.п. (а это может подкрепляться тем, что *A* способен выдавать и другие, столь же впечатляющие рассказы) – если это условие выполнено, тогда *A* можно, хотя бы условно, считать способным к творчеству. Любой компьютерный агент проходит ТЛ тогда и только тогда, когда он находится в том же положении относительно своего создателя, что *A* к Джонсу.

ТЛ опирается на основные эпистемические взаимоотношения, которые возникают между Джонсоном и *A*. Но «Джонс», как и «*A*», всего лишь некая переменная, обозначающая любого разработчика.

Это соответствует такому грубому, но эффективному определению ТЛ:

$\mathrm{Def}_{\mathrm{LT}}$: Искусственный агент *A*, созданный человеком *H*, проходит ТЛ если:

1. *A* создает *o*;
2. *o*, выданное *A*, не есть результат случайного стечения обстоятельств или сбоя машины, а результат, который *A* может повторить;
3. *H* (или кто-то иной, знающий тоже самое, что знает *H*, или обладающий способностями H^3) не может объяснить, как *A* выдал *o*.

Заметим, что ТЛ представим в форме мета-теста, так как схему этого теста можно использовать для любого частного случая. Например, если ТЛ использовать в контексте разговора, то *o* принимает вид предложения (или последовательности предложений, как и ТТ, конечно). Если ТЛ используется в контексте математических доказательств, то вместо *o* будет не предложение, а доказательство.

Возникают справедливые вопросы. Ниже представлены три из них, которые были заданы в ходе обсуждения первых версий статьи:

1. Какие ресурсы и знания имеет в своем распоряжении *H*;
2. Что можно считать удачным (правильным) объяснением;
3. Сколько времени есть у человека, чтобы дать это объяснение.

Ответ на третий вопрос прост: столько, сколько он потребует, исходя из здравого смысла. Предполагаемое объяснение не обязано прийти сразу: в распоряжении *H* месяц, месяцы, год, годы. Хотя сроки более двух лет можно считать бессмысленным. Мы понимаем, что эти временные параметры не совсем точные, но к ТЛ не надо применять стандарты ТТ и его различных вариантов[6]. Главное условие, очевидно, это то, чтобы *H* имел времени больше, нежели чем достаточно для того, чтобы во всём разобраться.

Первый и второй вопросы намного сложнее. Чтобы дать на них ответ, необходимо объяснить термин «искусственный агент», который является сердцевиной ИИ. К счастью, на пороге веков все лучшее об ИИ сконцентрировалось вокруг концепции, заслуживающей особого внимания – концепции интеллектуального агента (intelligent agent). Объединение ИИ-исследований произошло, в основном, благодаря всеобъемлющему руководства Расселла и Норвига (Russell and Norvig) (1994) («Искусственный интеллект: современный подход» – «*Artificial Intelligence: A Modern Approach (1994)* (AIMA). На обложке книги написано: «Книга об интеллектуальных агентах» («The Intelligent Agent Book»).

Полная, но без соблюдения формальностей структура интеллектуального агента показана на рис.1 (Взято прямо из текста AIMA). В соответствии с этой архитектурой, агент воспринимает перцепции из окружающей среды, обрабатывает их таким образом, чтобы выработать какие-

[3] Например, вместо *H* подойдет ученый, который наблюдал и «впитывал в себя» все то, что разработчики и сборщики *A* делали на каждом этапе

либо действия, совершает эти действия, получает новые перцепции и продолжает далее этот цикл.[4] В ТЛ действия искусственного агента заключаются в выводе результата, обозначенного переменной *o*.

В AIMA интеллектуальные агенты классифицируются: менее интеллектуальные, более интеллектуальные, самые интеллектуальные. Наименее интеллектуален «управляемый таблицей агент» (Table-Driven-Agent), программа в псевдокоде приведена на рис.2. Предполагается, что имеется набор действий, каждое из которых состоит в названии цвета («Зеленый», «Красный» и т.д.); так же предполагается, что перцепции - это числовые выражения цвета какого-то объекта, полученные с помощью сенсора этого агента. Дана Таблица 1 для нашего простого агента, выполняющего программу (рис.2). Так, агент произносит (например, с помощью синтезатора речи) слово «Голубой», если от сенсора идет 100.

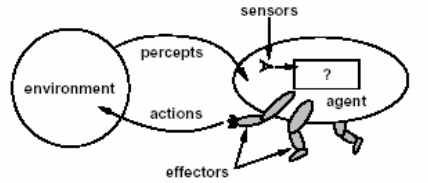

Рис 1. Структура Интеллектуального Агента

Конечно, это очень слабый агент. Любой выходной сигнал будет легко объяснён разработчиком. А как насчет агентов поумнее, которые хотя бы в некотором приближении могут стать кандидатами на прохождение ТЛ?

```
function TABLE-DRIVEN-AGENT(percept) returns action
  static: percepts, a sequence, initially empty
          table, a table, indexed by percept sequences, initially fully specified

  append percept to the end of percepts
  action ← LOOKUP(percepts, table)
  return action
```

Рис.2. Наименее интеллектуальный искусственный агент

[4] Цикл здесь поразительно схож с всеобъемлющей архитектурой познаний, описанной Поллоком (Pollock, 1995)

Таблица 1. Таблица поиска для агента, управляемого таблицей

Percept	Action
001	"Red"
010	"Green"
100	"Blue"
011	"Yellow"
111	"Black"

В AIMA мы найдём описание искусственных агенты, которые могут «удивить». Это – «основанные на знаниях агенты» («knowledge-based» agent). У них столь развитая структура, что они уже могут поспорить с ТЛ. Программа такого агента приведена на рис.3. Программа предполагает агента, у которого есть своя база знаний (*KB*), описывающая в формулах логики первого порядка следующие функции:

- TELL, которая репрезентирует факты в *KB*;
- MAKE-PERCEPT-SENTENCE, которая генерирует выражение по правилам логики первого порядка из перцепции и зависит от времени *t* (от времени восприятия перцепции);
- MAKE-ACTION-SENTENCE, которая генерирует повествовательное предложение (снова в форме исчисления предикатов) и выражающее действие, предпринятое в определённое время *t*.

```
function KB-AGENT(percept) returns an action
  static: KB, a knowledge base
          t, a counter, initially 0, indicating time

  TELL(KB, MAKE-PERCEPT-SENTENCE(percept, t))
  action ← ASK(KB, MAKE-ACTION-QUERY(t))
  TELL(KB, MAKE-ACTION-SENTENCE(action, t))
  t ← t + 1
  return action
```

Рис.3. Программа для «основанного на знаниях агента общего типа» (Generic Knowledge-Based Agent)

Эта небольшая консультация о природе искусственных интеллектуальных агентов позволяет нам, наконец, заметно продвинуться в ответах на первый и второй вопросы.

Итак, начнем.

1. Какие ресурсы и знания имеет в своем распоряжении *H*?

Ответ заключается в том, что предполагается, когда утверждается, что *H* обладает знаниями структуры агента. Это – знание *KB* агента и зна-

ние того, как осуществляются основные функции (например, как выполняются TELL и ASK) и т.п. Также предполагается, что *H* располагает достаточными ресурсами, чтобы точно определить эти элементы, «заморозить» и исследовать и пр. Однако это все не достаточно определённо. Чтобы прояснить, приведём пример. Этот пример имеет цель ответить на второй вопрос, который звучит так:

2. Что можно считать удачным /правильным/ объяснением?

Предположим, что на выходе от искусственного агента *A*' мы имеем, основанное на анализе, доказательство решения проблемы, с которой математики и логики безуспешно сражались десятилетия. Эта проблема, положим, заключается в определении: можно ли с помощью некоторой формулы φ вывести некоторый согласующийся набор постулатов Г. Представим, что после многих лет бесплодных размышлений некий человек *H*' кодирует Г и $\neg\varphi$, подаёт эти коды на вход OTTER-а и OTTER выдаёт доказательство, что данный способ кодирования дает противоречивый результат, т.е. $\Gamma \vdash \varphi$. Такой вывод приводит к массовым публикациям в СМИ по поводу «выдающихся» и «творческих» машин и т.д.[5] Однако в этом случае *A*' не проходит ТЛ в силу того, что *H*, зная *KB*, структуру и главные функции *A*', способен дать качественное объяснение доказательства – т.е. объяснить поведение агента *A*'. Объяснение не будет принципиально отличаться от доказательства OTTER-а, приведенного выше, но займет больше времени. Один из нас (Брингсйорд) регулярно даёт объяснения подобного рода. *KB* – это просто кодированная совокупность $\Gamma \cup \{\varphi\}$, структура доказательства состоит из поисковых алгоритмов, использованных OTTER-ом, и основных функций, заключающихся в правилах взаимодействия, на основе которых делаются выводы, используемых при доказательстве теорем. Если перевести всё это в термины Аргумента Китайской комнаты, то *A* не проходит ТЛ, потому что *H*, имея те же данные, что и машина, может вручную расставить символы в соответствии с этими знаниями и выдать искомое доказательство.

Ниже приведено вышеупомянутое определение в более совершенной форме:

Def_{LT} 2: Искусственный агент *A*, созданный *H*, проходит ТЛ тогда и только тогда, когда:

1. *A* создает *o*;
2. *o*, выданные *A*, не результат случайного стечения обстоятельств или сбоя машины, а результат, который *A* может повторить;
3. *H* (или кто-то, кто знает тоже самое, что и *H* и обладает возможностями H^5) не может объяснить, как *A* выдал *o* даже если он детально проанализировал структуру базы данных и изучил основные функции *A*.

[5] Для большей жизненности ситуации мы использовали решение OTTER-ом проблемы Робинсона (the Robbins Problem) см. в (Wos 1996).

4. Каких успехов добились современные системы в прохождении теста Лавлейс

Современные системы, специально ориентированные на реализацию творческих процессов или их имитацию однозначно проваливают ТЛ. Для всех этих систем разработчики легко могут применить АКК, т.е. они могут подставить себя на место машины, генерирующей ответы на вопросы простыми перестановками символов, в соответствии с информацией, представленной в базах знаний, алгоритмах и программных кодах. Осуществим анализ таких «творческих» систем:

4.1 BRUTUS

Сначала обратимся к системе BRUTUS (Брингсйорд и Ферруччи 2000). Система разработана с целью моделирования литературного творчества. Чтобы сформулировать задачу в духе ТТ, модно показать, что BRUTUS представляет собой результат многолетних попыток создать систему, способную имитировать «творение» коротких рассказов (the short short story game) или коротко тест S^3G (Bringsjord 1998*a*). (См. рис.5)

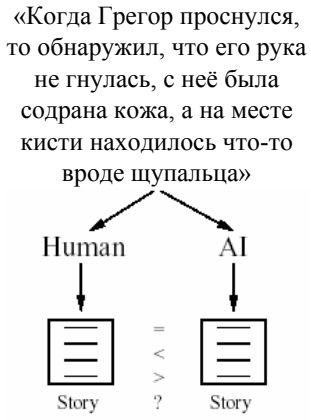

Рис.5. Игра в короткий рассказ

Идея, которую воплощает тест S^3G, проста. Человек и компьютер соревнуются. Оба получают одно, сравнительно простое, предложение, скажем: Однажды утром, проснувшись от тревожных снов, Грегор Самса понял, что его прямо в постели превратили в гигантское насекомое (Kafka 1948 стр.67). Оба, человек и машина, должны создать короткий рассказ

(порядка 500 слов) причём так, чтобы он был действительно интересным, имел литературные достоинства, был лучшим. Цель создания BRUTUS-а: создать искусственного писателя, способного соревноваться с лучшими писателями-людьми в S^3G, во многом как Deep Blue, идущий «голова в голову» с Каспаровым.

В чем же BRUTUS преуспел? Относительно цели – прохождения S^3G – не очень. Но с другой стороны, BRUTUS может «писать» довольно интересные истории, например:

«Предательство под влиянием самообмана»

Дэйв Страйвер любил свой университет. Он любил его увитые плющом башни с часами, древние мощные камни, залитые солнцем лужайки и активную молодежь. Так же ему нравился тот факт, что университет свободен от суровых, не прощающих ничего, испытаний из мира бизнеса – правда это не факт – в академии свои испытания, порой столь же беспощадные, что и на рынке. И главным примером служит защита диссертаций. Чтобы получить докторскую степень по философии нужно пройти устный экзамен на тему своей диссертации. Это была проверка, которую профессор Эдвард Харт любил устраивать.

Дэйв отчаянно хотел стать доктором, но ему нужны были подписи трех человек на первой странице диссертации, бесценные подписи, которые вместе подтвердят, что он прошел экзамен. Одна из подписей должна была прийти от профессора Харта, а Харт часто поговаривал – другим, да и самому себе – что для него – честь помочь Дэйву обрести заслуженную мечту.

Перед защитой Страйвер дал Харту предпоследний вариант диссертации, Харт прочел ее и сказал Дэйву, что она просто отличная и что он с радостью подпишет ее на защите. Они даже пожали друг другу руки в заставленном рядами книг офисе Харта. Дэйв заметил, что взгляд был светел и доверчив, а отношение было отеческим. На защите Дэйв думал, как великолепно все завершил третьей главой. Было два вопроса, один от профессора Родмана, другой от доктора Тира. Дэйв ответил на оба ко всеобщему удовлетворению. Дальнейших преград не было.

Профессор Родман подписал и плавно подвинул работу к Тире, она тоже поставила подпись, а затем положила ее перед Хартом. Харт не сделал ни единого движения.

«Эд!» - вопросительно произнес Родман.

Харт продолжал сидеть неподвижно. У Дэйва слегка закружилась голова.

«Эдвард, вы собираетесь ставить подпись?»

Позже Харт встретил в своем офисе огорченного неудачей Дэйва. Он думал, как помочь Дэйву обрести заслуженную мечту.

Заметьте, что слово «писать» перед этим рассказом было поставлено в кавычки. Почему? Причина проста, очевидна и снова возвращает нас к возражению Лавлейс: BRUTUS не сотворил эту историю. Он способен к её генерации, потому что два человека, Брингсйорд и Ферруччи, потратили годы на то, чтобы постичь и формализовать генеративную способность (generative capacity), достаточную чтобы выдать этот и другие подобные рассказы. Авторы могут воспроизвести путь формализации и так же, как и компьютер, выдать такой же рассказ. Данный метод известен как «инженерный анализ». Очевидно, что при подстановке в определение ТЛ BRUTUS вместо *A* и Брингсйорда и Ферруччи вместо *H*, BRUTUS провалит тест. Если перевести в термины Китайской комнаты, оба, Брингсйорд и Ферруччи, могли занять место компьютера, и бесмысленно манипулировать базой знаний, чтобы выдавать на выходе результаты, поражающие людей-судей.

Перейдем к рассмотрению частного случая, чтобы ещё более выпукло показать провал BRUTUS'а. BRUTUS запрограммирован сочинять необычные и странные истории, тем самым он привлекает читателей, которых захватывает нечто необычное и странное. В BRUTUS'е, чтобы выразить необычность, модификаторы (слова характеризующие объект) связаны с этим объектом по схеме, названной bizzaro_modifiers. Ознакомьтесь с примером такой схемы для фразы «истекающий кровью»:

```
instance bleeding is a bizzaro_modifier
    objects are {sun, plants, clothes, tombs, eyes}.
```

То, что Брингсйорд и Ферруччи назвали «литературно усиленной грамматикой» или LAG может быть ещё более усилено неестественностью, стимулирующее появление странных образов у читателя. LAG вместе с bizzaro_modifiers может быть использовано BRUTUS, чтобы сгенерировать следующее предложение.

Глаза Харта были как два истекающих кровью солнца.

- BizarreActionAnalogy → NP VP like ANP
- NP → noun_phrase
- ANP → modifier (isa bizzaro_modifier) noun (isa analog of NP)

Это предложение, выданное BRUTUS'ом, является результатом работы программиста. Такие предложения не являются результатом самостоятельного мышления BRUTUS.

4.2. COPYCAT

Дуглас Хофштадтер, как знают многие читатели, много лет думает над проблемой творчества и строит системы для изучения творческих процессов и подтверждения результатов своих раздумий. На какой конкретной форме творчества он останавливается? И какие из его систем наиболее продвинуты? Типичным представителем таких систем является **COPYCAT**, детально описанная в (Hofstadter 1995). Предполагается, что **COPYCAT** будет решать проблемы, создавая «творческие аналогии» (creative analogies).

Проблема 1. Положим, строка букв *abc* стала строкой *abd*; как вы подобным образом измените строку *jkl*?

Проблема 2. Положим, что строка букв *aabc*; превратится в *aabd*; как вы измените подобным образом строку букв *ijkk*?

Имитатор дает *ijl* как ответ на Проблему 1, а в процессе «рассматривает» варианты: *ijd* и *ijj*. Для Проблемы 2 программа выводит *ijll* и «рассматривает» варианты: ***ijkl***, ***jjkk***, ***hjkk***, ***jkkk***, ***ijkd***, ***ijdd***, ***ijkk*** и ***djkk***. Это хорошие ответы? Считать ли их творческими? Хофштадтер говорит «да» на оба вопроса. Но, кажется, он не замечает, что отвечает таким образом только потому, что **COPYCAT** был разработан, чтобы воспроизводить ответы, которые он сам (и многие другие люди) намерен дать. И по этой причине **COPYCAT** дает ответы подобного рода, потому что использует правила подстановки символов, такие как «заменить самый крайний справа символ на последующий». Но разве что-то иное предоставляют эти правила? Это – правила, о которых **COPYCAT** «осведомлен»?

Как свидетельство прихотливой природы **COPYCAT**, ознакомьтесь с фактом, что один из нас после ознакомления с проблемами, в первый раз дал ответы *ijj* и *ijkj* (заметьте, второй ответ **COPYCAT** даже не «рассматривался»). При этом мы считали так: правило, обеспечивающее получение результата, основано на рифме. В *abc* вторая и третья буква рифмуются. Когда строка заменяется на *abd* рифма сохраняется, в строке *ijk* вторая и третья буквы рифмуются. Чтобы выполнилось правило, необходимо заменить *k*, на другой символ, который бы рифмовался с *j*, второй буквой. Единственная возможность – это *j,* отсюда и было выдано *ijj*. То же самое правило, по очевидным причинам, приводит к *ijkj*.

То, что **COPYCAT** не пройдет ТЛ становится ясно, если перевести его правила подстановки в правила доказательства теорем. Очевидно, что правила подстановки, которые использовал Хофштадтер, можно достаточно легко выразить формулами логики первого порядка. Например, пусть *l* – это функция, отображающая три символа (c_1; c_2; c_3) в буквенную строку,

так что *l*(*a; b; c*) дает ***abc***. Теперь *s* – функция упорядочения для 26 строчных букв *{a, b, c, ..., z}*. Тогда правило замещения будет следующим:

$$\forall x \forall y \forall z (r(l(x, y, z)) = l(x, y, s(z))).$$

Легко зафиксировать замену букв с помощью подобных правил, а также установить приоритеты (это можно сделать и с помощью OTTER-а, обсуждавшегося выше). Вывод решения будет тогда состоять в выводе доказательств после того, как начальная строка (например, *ijk* в Проблеме 1), поступила на вход в систему. Очевидно, что разработчик доказываемой теоремы данной версии COPYCAT будет точно знать, почему система выдала конкретный результат или, формулируя в терминологии Серля, формула может быть использована в «книге правил» Китайской комнаты и один из нас, используя ее таким образом, сможет сидеть внутри этой комнаты, заставляя поверить людей снаружи в то, что система «самостоятельно творит».

4.3. LETTER SPIRIT

Ничего не изменится, когда мы ознакомимся (коротко) с LETTER SPIRIT-ом, другой системой из Hofstadter, 1995. Предметные области, в которых работают BRUTUS и COPYCAT – это тексты. В отличие от этих программ, LETTER SPIRIT работает с визуальными данными. LETTER SPIRIT получил данное название, потому что предназначен для распознавания букв (колонки на рис. 7) и стилей (ряды на рис. 7). Более точно, LETTER SPIRIT работает в рамках пространства символов латинского алфавита[6]. Вначале LETTER SPIRIT был задуман для создания новых шрифтов в рамках пространства латинского алфавита, но потом Хофштадтер (и работавший с ним вместе в этом проекте МакГрау (McGraw)) посчитали, что это слишком сложно. Они приняли решение, что вместо этого LETTER SPIRIT, получив на вход несколько «отсеянных» букв некоторого шрифта в качестве перцепции, на выходе даст оставшиеся буквы шрифта; кроме того они решили, что данные шрифты будут относиться к одному ограниченному классу «решеточных шрифтов» (“gridfonts”), как Хофштадтер и МакГрау их назвали (каждая буква такого решеточного шрифта создана выбором из 56 данных отрезков рис. 8).

Такое отступление, конечно, является печальным фактом, потому что похоже, что агент, способный создать свои собственные шрифты в неограниченном пространстве латинского алфавита, имел бы шансы пройти ТЛ. Но у LETTER SPIRIT таких шансов нет. Причина кроется в том, что Хофштадтер и МакГрау могут точно объяснить, как LETTER SPIRIT вы-

6 Часть этого пространства представлена на рис. 6. Согласно предположениям Хофштадтера пространство может быть несчётным (Hofstadter, 1982). Обсуждение этой проблемы см. в (Bringsjord & Zenzen 2001).

водит остальные буквы конкретного шрифта, обращаясь к базам знаний и алгоритмам. В терминах логики Сёрла, они оба по отсеянным буквам могут бессмысленно дописывать шрифты.[7]

В этом месте пора обратиться к возражению, которое, несомненно, многие наши читатели жаждут.

Figure 6: Various Letter As

Различные варианты написания буквы «А»

[7] Это вовсе не удивительно в силу метода, использованного Хофштадтером и МакГрау: основой для LETTER SPIRIT-а послужили знания о том, как человек создает и дополняет решеточные шрифты. Некоторые созданные людьми решеточные шрифты приведены на рис. 9

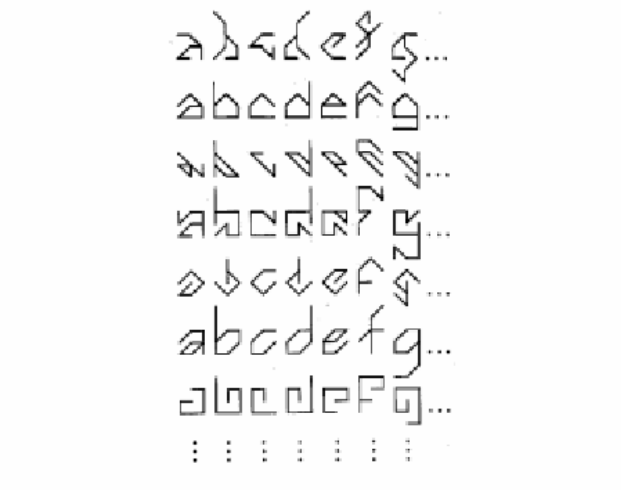

Figure 7: "Letter" is Covered by Columns, "Spirit" by Rows

Буквы расположены по колонкам, а стиль по строкам

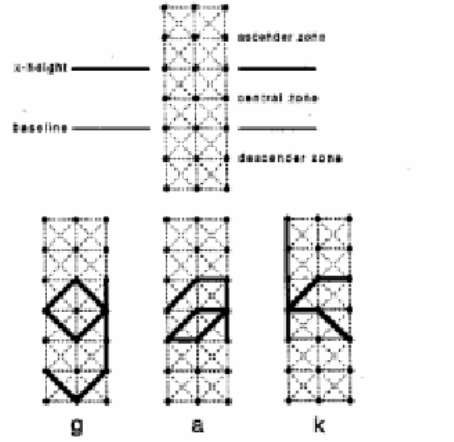

Figure 8: The 56 Quanta Defining Gridfonts

Решеточный шрифт на базе 56 составных отрезков

5. Возражение с позиции обучаемости

Возражение звучит следующим образом: «Господа, вам никогда не видать премии Тьюринга, потому что вы не заметили, что перед ответом на аргумент леди Лавлейс, имевший целью доказать, что компьютеры никогда не удивят нас, у Тьюринга ранее уже имелся пункт, состоящий из одного предложения, а именно: «Вопрос будет обсуждаться снова, но при изучении обучаемых машин (Turing 1964, p. 21). Говоря «про вопрос» Тьюринг ссылается не на что иное, как на вопрос, которым вы трое, следуя за Лавлейс, занимаетесь: может или нет компьютер создавать что-либо самостоятельное. Главное здесь – это «обучаемые машины». То, о чем говорит Тьюринг в 7 разделе своей статьи, разрушает протест Лавлейс подобно тому, как разрушает и ваш протест против способности машины создавать что-либо собственное, своё. На самом деле современные воплощения того, что Тьюринг предвидел в этом разделе, а именно, системы на нейронных сетях, которые по общему мнению, победят неприступный и суровый ТТ.

Позвольте объяснить.

«Стратегия Тьюринга при построении машины, способной пройти ТТ состоит не в программировании машины, которая способна жульничать на основе имеющихся знаний. Его стратегия, выраженная в разделе об обучаемых машинах, вместо этого заключается в создании «машины-ребенка» (child-machine), которую следует обучать так же, как и вас обучали в детстве. Вот цитата, которая должна задеть вас обоих».

> Важная черта обучаемых машин состоит в том, что их учитель часто не будет понимать, что происходит у них внутри, хотя он сможет в некоторой степени предсказывать поведение ученика. Наиболее сильно это проявится на поздней стадии обучения машины, возникающей из «машины-ребенка». Точка зрения, согласно которой «машина может делать только то, что мы знаем, как приказать ей это сделать» становится странной перед лицом данного обстоятельства (Turing 1964, p. 29)

Это возражение преодолевается довольно легко, как выяснится ниже. Предположительно, что машина-ребенок развивает свои способности посредством обучения искусственных нейронных сетей (artifucial neural networks) ANN. Теперь информация, обрабатываемая ANN, отвечает во всех ситуациях стандартным вычислением (например, алгоритму функционирования машины Тьюринга) или стандартной дедукции (например, доказательству в логике первого порядка) (см. Брингсйорд, 1991). Чтобы установить все точно, обратимся к понятию представимости, использованное в некоторых современных доказательствах теоремы Гёделя (Gödel):

некая функция f: $N^m \rightarrow N$ называется **представимой** в наборе Ф формул логики первого порядка, если есть формула $\varphi(v_1, \ldots, v_{m+1})$ (т.е. формула, свободные переменные которой $v_1, \ldots, v_{m+1}$) такие, что для всех $n_1, \ldots, n_{m+1} \in N$ выполняется:

- если $f(n_1, \ldots, n_m) = n_{m+1}$ тогда $\Phi \vdash \varphi(\tilde{n}_1, \ldots, \tilde{n}_{m+1})$
- если $f(n_1, \ldots, n_m) \neq n_{m+1}$ тогда $\Phi \vdash \neg \varphi(\tilde{n}_1, \ldots, \tilde{n}_{m+1})$
- $\Phi \vdash E^{=1} v_{m+1}\ \varphi(n_1, \ldots, n_{m+1}, v_{m+1})$

Мы говорим, что $\varphi(v_1, \ldots, v_{m+1})$ **представляет** функцию f в Ф.

Теперь положим, что машина-ребенок M_c в силу своих ANN вычисляет некоторую функцию f. Эта функция представима в некотором наборе Ф. Можете в этом случае считать, что Ф – это база знаний. Но тогда здесь больше уже нет никакого «самостоятельно мыслить», здесь командует специалист в области вычислительной техники и информатики, к него есть база занний Ф и правила дедукции. Основания этого ученого называть машину-ребенка марионеткой подобны основаниям, вынуждающими разработчиков BRUTUS-подобных систем признавать, что система ничего не творит.

6. Пройдут ли машины-оракулы Тест Лавлейс?

Какая же система способна пройти ТЛ? Как система может прорвать границы простой манипуляции символами и сделать что-то своё? Можно принять во внимание, что к этому способны так называемые «машины-оракулы». Так, Джек Копланд (Jack Copeland, 1998) недавно доказывал, что машины-оракулы (или как он их называл “О-машины”) защищены от утверждения Сёрла (1980), т.е. от АКК по поводу того, что простая манипуляция символами не достаточна для подлинного мышления. Системы проваливают ТЛ с того момента, когда становится ясно, что они перемещают символы предсказуемым образом. Идея Копланда заключалась в том, что машины-оракулы могут пройти ТЛ, потому что они делают нечто большее, чем манипулируют символами.

К сожалению, машины-оракулы не защищены от АКК. Характеристика Копланда, данная этим машинам, очевидно, неправильная. Машины всё-равно ничего не могут дать нового, что помогло бы в деле прохождения ТЛ.

Если кто-то и разработает «супер-голубя», который будет давать на выходе «супер-информацию», то у него возникнет убеждение в том, что эта птица самостоятельно думает, в противоположность медленному, «обычному голубю». Таким образом, лишь в силу убеждённости разработчика, «супер-голубь» выходит из ограничений леди Лавлейс.

7. Что же значит пройти тест Лавлейс?

Многие читатели, несомненно, спросят: Что же вы можете сказать *конструктивно*? Пока что все, что вы нам предъявили, не дало никаких результатов. Как насчет чего-нибудь практического? Что должен сделать сторонник ИИ, чтобы создать систему (способную пройти ТЛ), которая преодолеет границы бессмысленной манипуляции символами и начнет мыслить самостоятельно? Мы заканчиваем краткой теорией, поднятой этим вопросом.

К сожалению, нам кажется (по крайней мере одному из нас – Брингсйорду), что урок всего вышеизложенного совершенно отрицателен. Может просто не оказаться такого пути, по которому созданное человеком примитивное устройство, обрабатывающее информацию, пройдет ТЛ, так как то, что подразумевает Лавлейс, находится за пределами доступных нам причинно-следственных связей и математических отношений. Понятие о том, что способность к творчеству требует свободы и независимости, также признаёт и Хофштадтер, по крайней мере на стадии зарождения творческого процесса (например, см. pp. 411 в Hofstadter, 1995). Он уверен, что развитие компьютерной техники дойдет, в конце концов, до овладения такой формой свободы, которую используют творческие люди. Но что если способность «самостоятельно мыслить», требуемая ТЛ, влечет за собой такую форму свободы, как действия агента в соответствии со своей собственной мотивацией (*agent causation*). Теория мотивации агента (см. «главу VII: Свободная воля» в Bringsjord, 1992) влечет то, что люди воспроизводят привычный порядок вещей (например, осознают события или решения) напрямую без какой-либо причинной цепи событий. Хотя некоторые известные философы подтверждают такой взгляд (например, Ричард Тэйлор и Родерик Чизхолм (Richard Taylor and Roderick Chisholm)), их идеи сегодня не популярны.

Наша работа не имела намерения развивать идею «мотивации» машин. Ни одного аргумента в пользу этого взгляда здесь четко не сформулировано. Цель работы – обозначить трудность, которая появилась у нас во время создания концепции (не говоря уж о создании рабочего программного прототипа) искусственного агента, проходящего ТЛ. Эту трудность можно выразить в том факте, что творческий агент должен обладать довольно радикальной формой независимости и свободы. Говоря иначе, все разработки в области компьютерного творчества предоставляют индуктивные доказательства в пользу того, что мотивация реально имеет место в нашей созидательной, творческой, [собственно человеческой] жизни.

Литература

Boolos, G. S. & Je_rey, R. C. (1989), Computability and Logic, Cambridge University Press, Cambridge, UK.

Bringsjord, S. (1991), `Is the connectionist-logicist clash one of ai's wonderful red herrings?', Journal of Experimental & Theoretical AI 3.4, 319-349.

Bringsjord, S. (1992), What Robots Can and Can't Be, Kluwer, Dordrecht, The Netherlands.

Bringsjord, S. (1995), Could, how could we tell if, and why should-androids have inner lives?, in K. Ford, C. Glymour & P. Hayes, eds, `Android Epistemology', MIT Press, Cambridge, MA, pp. 93-122.

Bringsjord, S. (1998a), `Chess is too easy', Technology Review 101(2), 23{28.

Bringsjord, S. (1998b), Philosophy and `super' computation, in J. Moor & T. Bynam, eds, `The Digital Phoenix: How Computers are Changing Philosophy', Blackwell, Oxford, UK, pp. 231-252.

Bringsjord, S. & Ferrucci, D. (2000), Arti_cial Intelligence and Literary Creativity: Inside the Mind of Brutus, a Storytelling Machine, Lawrence Erlbaum, Mahwah, NJ.

Bringsjord, S. & Zenzen, M. (2001), SuperMinds: A Defense of Uncomputable Cognition, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Copeland, B. J. (1998), `Turing's O-machines, searle, penrose and the brain', Analysis 58(2), 128-138.

Ebbinghaus, H. D., Flum, J. & Thomas, W. (1984), Mathematical Logic, Springer-Verlag, New York, NY.

Gold, M. (1994), `Limiting recursion', Journal of Symbolic Logic 30(1), 28-47.

Harnad, S. (1991), `Other bodies, other minds: A machine incarnation of an old philosophical problem', Minds and Machines 1.1, 43-54. This paper is available online at ftp://cogsci.ecs.soton.ac.uk/pub/harnad/Harnad/harnad91.otherminds.

Hofstadter, D. (1982), `Metafont, metamathematics, and metaphysics', Visible Language 14(4), 309-338.

Hofstadter, D. (1995), Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought, Basic Books, New York, NY.

Hofstadter, D. & McGraw, G. (1995), Letter spirit: Esthetic perception and creative play in the rich microcosm of the roman alphabet, in `Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought', Basic Books, New York, NY, pp. 407-488.

Kafka, F. (1948), The metamorphosis, in F. Kafka, t. W. Muir & E. Muir, eds, `The Penal Colony', Schocken Books, New York, NY.

Kugel, P. (1986), `Thinking may be more than computing', Cognition 18, 128-149

Moor, J. H. (1976), `An analysis of turing's test', Philosophical Studies 30, 249-257

Moravec, H. (1999), Robot: Mere Machine to Transcendant Mind, Oxford University Press, Oxford, UK

Pollock, J. (1995), Cognitive Carpentry: A Blueprint for How to Build a Person, MIT Press, Cambridge, MA

Putnam, H. (1994), `Trial and error predicates and a solution to a problem of mostowski', Journal of Symbolic Logic 30(1), 49-57

Russell, B. (1936), `The limits of empiricism', Proceedings of the Aristotelian Society 36, 131-150.

Russell, S. & Norvig, P. (1994), Arti_cial Intelligence: A Modern Approach, Prentice Hall, Saddle River, NJ.

Searle, J. (1980), `Minds, brains and programs', Behavioral and Brain Sciences 3, 417-424. This paper is available online at http://members.aol.com/NeoNoetics/MindsBrainsPrograms.html.

Siegelmann, H. (1995), `Computation beyond the turing limit', Science 268, 545-548.

Siegelmann, H. & Sontag, E. (1994), `Analog computation via neural nets', Theoretical Computer Science 131, 331-360.

Turing, A. (1964), Computing machinery and intelligence, in A. R. Anderson, ed., `Minds and Machines', Prentice-Hall, Englewood Cli_s, NJ, pp. 4-30.

Weyl, H. (1949), Philosophy of Mathematics and Natural Science, Princeton University Press, Princeton, NJ.

Wos, L. (1996), The Automation of Reasoning: An Experimenter's Notebook with otter Tutorial, Academic Press, San Diego, CA.

Арон Сломэн

Что значит быть камнем?

(университет Бирмингема, Англия, Великобритания)

Перевод выполнен по работе**: Sloman Aaron, 1996. WHAT IS IT LIKE TO BE A ROCK?**
(http://www.cs.bham.ac.uk/~axs/misc/like_to_be_a_rock/rock.html)

Перевод Василия Крючкова

Краткий обзор содержания данной статьи

В контексте нападок на концепцию сильного искусственного интеллекта возникает множество псевдовопросов. Отвечать на них бесполезно, это – пустая трата времени. Данная статья призвана заменить псевдовопросы настоящими вопросами, обусловленными выяснением природы и сущности внутренних состояний разумных существ самых различных видов. В частности, будет изучаться вопрос «Что значит быть X?», который обычно звучит при объяснении феноменов, для которых недостаточно физических утверждений и которые нельзя воспроизвести в компьютерных реализациях агентов (существ). В работе данный вопрос рассматривается с двух позиций: (а) с позиции определений, которые, хотя и неполно, характеризуют состояния и способности рассматриваемого агента. Это позволяет задать онтологию, необходимую для моделирования данного агента; (б) с позиции бессмысленных, непонятных определений.

Работа придерживается анти-редукционистской точки зрения, нет никакого дуализма и мистики.

Предисловие

Дискуссии по проблемам сознания и самосознания, включающие также вопросы возможности сознания у машин (т.е. можно ли, в принципе, наделить машину сознанием), по проблеме того, каких животных следует считать сознательными, по проблеме формирования, развития и функционирования сознания – дискуссии по этим и другим проблемам обычно основаны на нашем высокомерном предположении, что всем известно, что означает слово «сознание». Кроме этого, предполагается, что всё, что имеется в ввиду под «сознанием», можно определить множеством общеупотребительных, общепринятых выражений, т.е. элементарных выражений, которые получили статус специальных терминов и которое можно использовать в философских и научных дискуссиях. В дискуссиях наиболее часто обсуждается такая форма постановки проблемы «Что значит быть X?»

(Т. Нагель, 1981). Дискуссии в такой форме обычно используются на comp.ai.philosophy.newsgroup.

Лично я убеждён, что большинство дискуссий о проблемах сознания основывается на высокомерном представлении о прозрачности ставящихся вопросов. Поэтому цель данной статьи в том, чтобы немного поубавить пыл. Я не отрицаю существование ряда вопросов, которые следует ставить в контексте проблемы сознания. Однако я утверждаю, что нужен новый подход к самой постановке этих вопросов, основанный на группировании различных видов интеллектуальных агентов и анализе типов состояний и процессов, характерных для каждой такой группы.

К сожалению, соблазн поставить всё те же псевдовопросы очень силён. Поэтому целью данной работы является также сведение этого соблазн на «нет». Но примут это не все читатели.

Подход к дискуссии

Мой подход состоит в более-менее подробном обсуждении следующего вопроса:

Что значит интересоваться тем, что значит быть X?

Я постараюсь проанализировать, что нужно для того, чтобы дать удовлетворительные ответы, начиная от нескольких частных точек зрения и завершая философской позицией, которая является как антиредукционистской, так и функционалистской.

Что значит быть камнем, одним из многих, лежащих повсюду?

Я, конечно, не знаю исчерпывающего ответа на этот вопрос, однако, в отличие от самого камня, знаю многое.

Это значит быть размером около фута в диаметре.

Это значит весить несколько фунтов.

Это значит в основном, состоять из кремния (я так думаю).

Это значит покоиться на пологом склоне, на небольшом грязном клочке земли.

Иногда это значит быть отброшенным в сторону, подброшенным в небо и упавшим с гулким звуком удара о землю.

И, наконец, быть камнем, значит никогда не знать ничего из вышеперечисленного.

Некоторые возразят мне, что я искажаю суть вопроса и далее решаю не проблему «Что, значит, быть X?», а проблему «На что похоже X?». Но раз мы хотим увидеть мир через восприятие X, то здесь предполагается, что у X есть точка зрения, следовательно, необходимо давать объяснение с точки зрения X, несмотря на то, что впоследствии уже вовсе не потребуется, чтобы у X была своя точка зрения. Проблема, конечно, остаётся прежней, она только видоизменяется при переводе её в новую плоскость: в

плоскость изучения того, что такое «точка зрения» и какие вещи ею «обладают». Есть ли у подсолнуха «точка зрения»?

Что значит быть подсолнухом?

Здесь я уже не могу описать всё так полно, как хотя бы в случае с камнем.

Это значит иметь возможность расти, распуская листья и корни для обеспечения себя питанием и опорой.

Но это значит не иметь возможность ходить или бегать.

Это значит обладать информацией о положении солнца на небосводе, но, в то же время, не иметь информации о многом другом, что творится в окружающем мире.

И многое ли может сделать подсолнух с доступной ему информацией? К примеру, он не может по положению солнца определить время суток и пора ли детям идти в школу.

Не похоже, что подсолнух много знает о таких вещах как верх или низ, солнце и луна или знает, почему ему полезнее всего поворачиваться к солнцу.

Он имеет точку зрения, но в том смысле, что точка зрения – это такое положение в мире, которое обеспечивает информацию о нём. С разных точек зрения, зачастую, доступна разная информация. Но вопрос заключается не только в раскрытии физических или геометрических понятий «точки зрения», но ещё и в том, какой информацией подсолнух (или любое другое существо) действительно располагает и как он её использует. Это уже вопрос к биологам. Поиск решения может быть очень сложным, наши знания по этому поводу невелики и разрознены. Но из неполноты этих знаний не следует некая особая тайна.

Что значит быть летучей мышью?

Это, очевидно, гораздо веселее всего рассмотренного ранее. Это значит парить на высокой скорости, даже в темноте.

Наверное, это должно немного напоминать состояние летящего на гребне волны сёрфингиста. Но летучая мышь контролирует полёт гораздо лучше.

Это, в некоторой степени, значит уметь издавать и слышать звуки, слишком высокие, для того, чтобы их мог услышать человек. Но это, однако, далеко не значит иметь возможность спеть партию сопрано из концерта Моцарта или просто насладится, слушая этот концерт.

Это значит иметь возможность на слух определять расстояние до окружающих объектов. Но это не значит с помощью хлопка определять, велика ли комната.

Иногда это значит очень долго висеть вверх тормашками, не испытывая дискомфорта.

Мы можем сказать о том, что значит быть летучей мышью гораздо больше, чем о том, что значит быть подсолнухом – летучая мышь выполняет много больше разнообразных функций, и, по сравнению с ними, она гораздо больше похожа на нас с вами. Это позволяет распространять на неё и использовать для описания её мира наши категории.

С другой стороны, что значит накапливать знания или использовать информацию, как это делает летучая мышь? Это те вещи, которые нам не дано понять в деталях, т.к. летучие мыши, насколько я знаю, не используют при этом ничего похожего на наш мыслительный аппарат. До этого пункта дошёл Нагель в своей работе. Но далее он не предлагает никаких объяснений.

А объяснение может быть таковым – в процессе обработки информации механизмы, используемые летучими мышами, крайне отличаются от наших информационных механизмов. Более того, те процессы, которые происходят у них и те состояния, в которых они находятся, невозможно спроецировать на наше сознание без серьёзных потерь.

Здесь, конечно, можно провести некоторые параллели. Например, у нас имеются общие возможности по постижению пространства, времени и движения. Однако обнаружатся значительные расхождения при построении систем связей между объектами, например, при определении сексуальной привлекательности партнёра или при выборе продуктов питания по вкусу.

Их мир сильно отличается от нашего. Нет причин верить в то, что у летучих мышей категоризация вещей, состояний, событий, процессов, действий или чего-либо замещающего всё это при их способе мышления, может быть прямо перенесено в категоризацию, которой пользуемся мы.

Вкратце, перефразируя Витгенштейна: если бы летучая мышь могла бы сказать нам, что значит быть летучей мышью, мы просто не смогли бы её понять.

Имеется, по крайней мере, две причины этому: во-первых, у нас сильно отличаются потребности и способы их реализации в окружающем мире (так, чтобы утолить голод, мы не летаем, охотясь за насекомыми) и, во-вторых, наш способ восприятия информации может отличаться, пускай даже незначительно. Например, нам, возможно, не дано испытывать одинаковые эмоциональные состояния или мы не имеем такого же многообразия функционально различающихся, сосуществующих и взаимодействующих компонентов.

Репрезентационная грамматика у мышей иная, потому и семантические способности отличаются.

Некоторые языки невозможно выучить, не обладая необходимыми средствами для этого. Очевидно, что более простой способ отображения информации не справится с отображением состояний, более сложных в семантическом плане. Отображение информации будет в этом случае неполным. Тем не менее, всё то, что отобразится, будет истинным.

Конечно, невозможно «перевести» информационные структуры летучих мышей в наши. Однако, в принципе, нет ничего, что могло бы остановить на пути получения подробного объективного описания этих структур. Например, стало известно, что у летучие мыши издают 17 информативных сигналов и то, что их при получении информации от других мышей, они различают 37 элементов и используют 5 типов связей между этими элементами.

Полная теория летуче-мышиной семантики может потребовать расширения нашего логического метаязыка, ведь мышление летучих мышей может и не использовать понятия и связи между ними, а представлять лишь совокупность реакций на раздражители в условиях конкретной ситуации.

Конечно, изучение таких вопросов – сверхсложная задача. Более того, даже дав частичный ответ на вопрос «Что значит быть летучей мышью?» это всё равно не позволит нам прочувствовать само состояние летучей мыши. Нельзя это состояние познать непосредственно, нельзя даже представить его себе.

Важно, что такие описания не обязательно должны быть изложены на языке физики. Более вероятно, они должны быть представлены в терминологии лингвистов и специалистов в области компьютерных наук – эти специалисты утверждают о структурных особенностях информации, не вдаваясь в способ её физической реализации.

Что значит быть (нормальным) человеческим младенцем?

Это значит быть одновременно слишком слабым, слишком неслаженным и совершенно неосведомлённым о том, чтобы, к примеру, самостоятельно приготовить себе ужин.

Но это не значить, что всё это известно ребёнку, он, вероятно, ничего и не знает о своей слабости, координации, информированности или об ужине.

Быть младенцем, это, отчасти, видеть, слышать, чувствовать голод и боль. Однако вспомогательные мыслительные процессы, которыми пользуются младенцы и не пользуются взрослые люди описать невозможно. Если новорождённый заговорил бы с нами, то мы смогли бы понять из разговора с ним не много больше, чем из разговора с летучей мышью (если, к тому же родители нам разрешат поговорить с ребёнком).

Замечательным свойством человеческого разума является его развитие. Многие изменения можно заметить лишь со временем, однако куму-

лятивные эффекты могут приводить к существенным качественным изменениям [в системе знаний], например, к возникновению новых информационных структур (к примеру, знаний по квантовой механике) и новых управляющих архитектур (к примеру, к способам оттягивания уплаты чаевых). Мы очень мало знаем о подобных усовершенствованиях, неизвестны, например, какую роль в этих изменениях играет генетика, а какую взаимодействие со средой. Разумеется, среда, в которой происходит развитие ребёнка, играет очень важную роль, так через своё окружение ребёнок учится говорить на китайском, читать по нотам или понимать квантовую физику.

Поэтому у взрослого человека столь большое разнообразие информационных структур, что перевод в них знаний ребёнка невозможен. Таким образом, разница между тем, что я знаю о том, что значит быть летучей мышью и тем, что значит быть ребёнком, это только различие в степени развития [знаний].

Однако не всякое развитие – это рост или совершенствование, увы.

Что значит быть больным на одной из последних стадий болезни Альцгеймера?

Что значит быть подобного рода больным – об этом я лично знаю очень немного, всё то, что почерпнул из грустных и трогательных документальных телевизионных передач.

К примеру, в некоторых случаях это значит хотеть вытереть подоконник, не помня, что уже вытер его минуту или две тому назад.

Что, значит, быть умственно отсталым человеком?

На днях (в ноябре или декабре 1995) я смотрел вторую часть увлекательной телевизионной программы (Великобритания, ТВ4, кажется). В ней шла речь об умственно отсталой женщине, которая, несмотря на это, могла ясно выражать свои мысли и даже написала книгу. Ясно, что быть умственно отсталым, значит быть таким как она, т.е. в чём-то отличаться от того, что значит быть нормальной личностью.

Сломан (Sloman, 1989) цитируя (Self 1977), выдвинул гипотезу по поводу впечатляющих способностей к рисованию у умственной отсталой девочки Нади. Эти способности могли быть результатом неправильно завышенной чувствительности к информации, получаемой на нижнем, аналитическом уровне интерпретации видимой картины мира. У Нади по какой-то причине не функционировал высший уровень – синтетический, а этот уровень интегрирует интерпретационные механизмы и доминирует при построении нормальной картины мира.

Если так, то процессы, вырабатываемые нормальным мозгом, приводят к минимизации и даже вытеснению низкого уровня анализа и интерпретации с целью повысить значимость роли, которую играет уровень распознавания образов и рассудочного мышления. Эти процессы в аутичных мозгах не срабатывают. Здесь доминирует низший уровень обработки информации и ограничен или полностью отсутствует высокий, более абстрактный и холистичный уровень

Так, одна умственно отсталая женщина пыталась обрисовать свой поход в супермаркет. Специальные эффекты киносъёмки давали зрителям возможность почувствовать, на что похоже её видение мира: масса мелких образов, быстро меняющиеся детали, которые очень трудно объединить в общую картину происходящего. Конечно, и нормальный человек, при желании, может получить такой вид окружающей действительности, как тот, который я только что описал, но это всё равно будет не то же самое, что и получение комплексного взгляда на вещи. Никакие эффекты съёмки до конца не передают, что на самом деле значит быть умственно отсталым. Киносъёмка лишь *изменяет* высокий уровень синоптической характеризации, не подавляя его.

Если где-то мы и можем преуспеть в способах описания природы чуждых нам форм опыта, мы пока не способны не то, что владеть этим опытом – вообразить его не можем. Дескрипция всегда проще репликации.

Что значит быть слепым?

Мне не нужно быть Х, чтобы знать, что это значит. Слепой от рождения человек может многое знать о том, что значит видеть.

Быть зрячим – это значит получать информацию о поверхностях вещей – их текстуре, ориентации, положении в пространстве и при этом не испытывать необходимости прикасаться к этим вещам. Быть зрячим – это более эффективно воспринимать слона, нежели чем последовательно изучать его на ощупь.

Большая часть информации, полученная о предмете посредством визуальных и тактильных ощущений одинакова, так что различия в полученной информации не явны и проявляются в том, что используются для разных целей. Зрение, например, используется для быстрого сравнения двух поверхностей или для того, чтобы различать по внешнему виду других людей. Однако, если бы я, играя на скрипке, мог бы видеть свою руку, но не чувствовать её при этом, то играл бы хуже обычного.

Различать цвета, видеть их – это определять текстуру поверхности – цветовые и текстурные области, границы, оттенки, композиции, 2D-формы, такие как буквы или многоугольники, состоящие из областей одной текстуры или цвета и тд.

Видеть цвета, это, отчасти, означает получать информацию о всей поверхности очень быстро, на расстоянии, кроме того, получать её в пространственно структурированном виде (и не так долго, как читать это предложение).

Что теряется, когда человек лишён такого способа познания? Очень многое, я мог бы перечислить всё это в терминах, понятных зрячему человеку, но, вероятно, слепой бы не понял.

С одной стороны, следует помнить, что человек с врождённой слепотой и чьи глаза никогда не видели, сохраняет множество мозговых функций, которые используются нами для определения цвета и которые развивались на всём протяжении человеческой эволюции. Возможно, такой человек может, слушая, как зрячие люди разговаривают о цветоощущении и прочих вещах, связанных со зрением, воссоздать образы, сходные с теми, которые возникают в мозгу зрячих во время разговора. Здесь важен эмпирический вопрос – что именно слепой от рождения человек способен понять о зрении? А это зависит от того, какие возможности обработки информации заложены в его мозге.

Незрячие люди могут создавать картины на ощупь, они представляют собой топологию образов, которые рисуют и нормальные люди. Но углы, пропорции на картинах слепых нарушаются. Хотя, в принципе, такие нарушения наблюдаются и на рисунках зрячих маленьких детей.

Быть слепым, это, возможно, значит быть роботом, способ восприятия мира которого можно охарактеризовать следующим: видеосистема у него разработана для получения информации о трёхмерной структуре окружающего мира, в то время как двумерная структура оптической информации выступает в форме частного образа. Тот факт, что двумерная структура используется низкоуровневыми процедурами обработки зрительной информации, не означает, что робот может сознательно получать к ней доступ и использовать эту информацию для создания самостоятельного образа. Проблема состоит в создании цельной трёхмерной структуры из набора двухмерных образов, не прибегая при этом к знаниям о самих двухмерных образах, точно так же как это делаем мы, зрячие.

Кроме того, наличие возможности видеть что-то, ещё не означает возможности отображения виденного образа в рисунке. Для видения образа и рисования его требуется разная информация. Ведь, возможность видеть не подразумевает возможность позднее передать увиденное, воссоздав его на бумаге или ещё на чём-нибудь. Когда мы запоминаем увиденное нами в «обычном» режиме, из памяти восстанавливается только малая толика виденного (однако есть люди, например, вышеупомянутая Надя, которые могут восстановить по памяти больше деталей, чем другие).

Что значит быть женщиной?

Несомненно, что очень многое из того, что значит быть женщиной, я никогда не узнаю. Я знаю, что значит находить женщину привлекательной и, основываясь на этом, я могу по аналогии немного представить, что значит находить привлекательным мужчину, но далеко не до конца. Поэтому,

возможно, я больше знаю, что значит быть лесбиянкой, нежели чем гетеросексуальной женщиной.

Непосредственно я не знаю, что, значит, быть «продолжательницей рода», то есть, отчаянно желать родить ребёнка, но я знаю довольно-таки много из разговоров об этом. Я знаю, что значит быть сознательно или бессознательно подвергаться дискриминации, унижению или покровительству в этом «мужском» мире, хотя личного опыта я, конечно, не имею.

Кроме общих отличий могут быть ещё глубокие различия между теми, кто мыслит нормально, с одной стороны, и, с другой стороны, между умственно отсталыми людьми, у которых рассудок повреждён или затуманен старческими годами. Эти отличия определяются генетическими различиями в механизмах переработки информации.

Так ли это или нет и можно ли генами объяснить феномен редкой музыкальной, актёрской или научной гениальности, встречающийся среди людей? Ясно то, что существуют различия, обусловленные полом и которые несут положительную функцию – без генетических комплексов, составляющих человеческую суть, люди менее успешно воспроизводили бы себя.

Таким образом, разумный робот может быть бесполым – быть одновременно и как мужчина и как женщина.

Что значит быть роботом?

Для начала, «что значит быть роботом?», зависит от типа робота.

Робот, который обладает частичными возможностями человеческого восприятия мира, никогда не сможет до конца прочувствовать, что значит быть сёрфингистом, мчащимся на гребне волны; не сможет ощутить на себе действие центробежных и гравитационных сил; не почувствует дуновение ветра, слегка развевающее волосы; не услышит крик и не почувствует, как напуган и взволнован ребёнок, цепляющийся за вас; не испытает сухость во рту от собственного страха.

Но зато робот может многое узнать, например, узнать о том, что значит быть сёрфингистом, если робот будет способен к анализу информации о ситуации, в которой пребывает сёрфингист. Робот может многое знать о возможностях человеческого восприятия, мотивах, эмоциях. Другими словами, в него можно вложить колоссальные знания о человеческих способностях, даже если воспроизвести все эти знания робот полностью не сможет.

Итак, ответ на вопрос, что значит быть сложным роботом, который конструктивно очень сильно отличается от нас, включает в себя ответ на вопрос что *значит быть человеком*. Робот может знать лучше и больше о том, что значит быть человеком, чем о том, что значит быть летучей мышью. Это обусловлено причинами ограниченности наших знаний о летучих мышах в силу: а) недостатка информации; б) неподготовленности на-

шего категориально-понятийного аппарата для трансляции в него летуче-мышиной системы восприятия информации.

Часто утверждается (Нагелем и многими другими), что робот может симулировать способы поведения людей без всякого осознания того, что значит быть человеком. Точно так же он может симулировать поведение летучей мыши.

Но здесь игнорируется тот факт, что робот не может нормально симулировать, не обладая сложными механизмами обработки информации (Sloman 1994). Можно задаться вопросом – какого рода информацию получает робот о различных окружающих его объектах, о самом себе или своих внутренних структурах (McCarthy 1995).

Некоторые внутренние состояния робота невозможно воспроизвести в нашем мозгу, как и в случае с летучей мышью. Но мы можем описать их – то есть описать их инфорамационную структуру, семантику, трансформации и способы использования.

Необоснованно считать, что робот, функционирующий подобным образом, может быть нечто вроде «зомби», в том смысле этого понятия, что он будет повторять какие-то действия, без каких бы то ни было чувств, точек зрения или чего-то подобного. Те, кто утверждает, что могут представить подобное, занимаются самообманом. Они сродни тем, кто считает, что может придумать метод точного измерения трисекции любого угла, пользуясь лишь линейкой и транспортиром. Им следует поучиться инженерному делу, лишь тогда они поймут, что их робот-зомби работать не будет.

Для хорошего инженера представить себе робота-зомби означает отказаться от всего, что он знает о механизмах, которые позволяют роботу обрабатывать воспринимаемую информацию, генерировать новые задачи, выбрать между задачами с целью реализации поведения и воздерживаться от недопустимых действий, оценивать степень выполненности задачи или справляться с неожиданно возникшим внешним повреждением.

Я твёрдо стою на позиции необходимости выработки онтологии высокоуровневой обработки информации для любой системы, способы функционирования которой построены по аналогии с человеческим мышлением. Но здесь следует быть осторожным: теоретически допустимо иметь огромную справочную таблицу, в которой записаны все возможные варианты сенсорных входных данных, сохраняемые в памяти и которым приписываются внешние отклики. Если это физически возможно (а это не так для существ, которые обладают способностями, схожими с человеческими), тогда такой машине может и не потребоваться человекоподобная внутренняя онтология. Но в данном случае она была бы больше похожа на камень, нежели чем на летучую мышь. И если вы хотите представить себе зомби, то создайте такого робота, который на основании вышеприведённой таблицы мог бы функционировать как сознательное существо.

Предварительные выводы

Я попытался показать, рассматривая различные случаи, что нам многое известно о том, что значит быть Х, а ещё больше – о том, что мы не знаем в силу различного рода причин для каждого конкретного случая. Однако некоторые аспекты того, что значит быть Х, не всегда можно идентично перенести на У. Это случай, когда у У информационная система отличается от системы Х. Когда правильно разобрался с данным вопросом, понимаешь, что он относится к сфере теории информации, а не к философии!

Некоторые философские возражения

Многие философы согласятся со всем этим. Но некоторые из них (и многие из которых далеки от философии) будут протестовать, что ничего из вышеизложенного не приближает нас к пониманию, что, значит, быть Х «изнутри».

Я, конечно, могу рассуждать на тему, что значит быть Вами, говоря, что это означает: занимать Ваше местоположение, видеть вещи, которые можете видеть только Вы, чувствовать то, что не могу почувствовать я (так как Вы, например, сидите в удобном кресле, а я сижу на корточках на полу) и знать Ваш образ мыслей, то о чём я, в лучшем случае, могу лишь догадаться.

Если «быть Вами “изнутри”« означает что-то вроде обретения Вашего видения мира, включая в него Ваше сознание и тело в конкретный момент времени, тогда я, конечно, не могу точно знать, что значит «быть Вами “изнутри”« - для этого мне не достаточно информации.

Некоторую толику этих знаний я могу добыть. Например, я могу предположить, что Вы видите часть стены, которую от меня скрывает огромная колона. Я так же могу узнать, какое из моих высказываний разозлит вас, а какое заставит издеваться надо мной. Вы знаете это, а я – нет, но я знаю *о том, что вы знаете*.

Здесь нет большой разницы между а) знанием о том, что есть что-то вокруг вас, то к чему у вас есть доступ, а у меня нет и б) знанием о том, что есть что-то внутри вас, к чему вы имеете доступ, а я нет. Оба примера иллюстрируют разницу в доступе к информации.

Конечно, я могу устранить некоторые различия между нами, выйдя к Вам из-за колоны и увидев ранее скрытую от меня часть стены.

Но я не смогу найти такое место, где бы смог «увидеть» с Вашей точки зрения Ваше же внутреннее эмоциональное состояние.

Более того, если вы дальтоник, а я нет, или, наоборот, у меня не будет точного совпадения с вами даже при простом просмотре изобразительной картины. Может быть, придет время и изобретут пару шлемов, связанных оптоволокном, с помощью которых можно будет преодолеть между нами преграду и разделить с Вами Вашу точку зрения, так как если бы я

очень близко с Вами. Хотя для этого, наверное, можно использовать специально расположенные зеркала и камеры, что тоже поможет устранить преграду между нами, правда, в меньшей степени, чем с использованием воображаемых шлемов. Такие шлемы возможны и это несомненный будущий эмпирический факт. В «шлемах» нет ничего глубоко философского (по сравнению, например, с трудностью познания того, что происходит внутри субатомных частиц и вообще, существует ли в микромире это «внутри»).

Есть ли у компьютеров точка зрения?

К сегодняшнему дню создано большое количество сильно отличающихся друг от друга компьютерных систем. В них мы не найдем ничего схожего с нашими информационными системами – например, средства долгого хранения информации крайне сильно отличаются от нашей памяти (способы получения, доступа, сохранения и забывания информации совершенно отличны). Они (в основном) ничуть не нуждаются ни в чём похожем на человеческие мотивации, а те которые имеют возможности анализа задачи, в основном, получают их от кого-то ещё, поскольку человеческая сущность включает в себя разнообразные независимые источники мотиваций которые (в основном) не содействуют чужим целям. (Человеческие мотивы не всегда детерминированы индивидуальными целями и потребностями, многие вырабатывались в ходе длительной эволюции и преследовали цель сохранения и распространения генофонда. Наши гены посмеялись бы над нами, если бы могли смеяться, за те поступки, которые они заставляют нас совершать).

Из-за отличий между компьютерными системами и людьми (или другими животными) наше знание о том, что значит быть компьютерной системой с разделением времени не многим отличается от нашего знания, что значит быть камнем: в основном, это информация «от третьего лица» о том, что происходит, хотя, в отличие от примера с камнем, «быть компьютером» менее скучно.

Так или иначе, по мере усложнения систем искусственного интеллекта, развития у них самостоятельности как в механизмах мотивации, так и в развитии понятийного аппарата, который контролирует процессы восприятия и задаёт семантику для внутренних источников информации, будет возникать всё больше и больше вопросов на тему, как такая система должна классифицировать вещи и определять при этом контекст ситуации? Что значит быть такой системой с её «точки зрения»? Например, что она захочет сделать в зависимости от ситуации, с какими решениями ей придётся столкнуться, как она решит, что такое «сложно», а что такое «легко», что ей «нравится», а что «не нравится»?

Интеллектуальные системы будут всё больше и больше знать о том, что значит «быть роботом “изнутри”«.

Чтобы это не значило.

Конечно, данный процесс займёт уйму времени – может быть, сотни лет, прежде чем компьютеры выйдут на уровень шимпанзе и столько же времени до достижения ими уровня человека.

И когда это случится, некоторые роботы зададутся вопросом, что значит быть личностью или что значит быть летучей мышью. Или, может быть, что значит быть камнем.

Ещё немного философских возражений

«Но, но, но» - лопочет разочарованный оппонент, всё более раздражаясь из-за меня. «Вы до сих пор и близко не подошли к тому, о чём я говорил: что значит ПО-НАСТОЯЩЕМУ «ИЗНУТРИ» быть летучей мышью или личностью. Всё то, о чём вы до сих пор говорили, отвечает на вопрос: что значит иметь различные физические структуры и процессы, а также иметь некоторые нефизические способности к осуществлению процессов обработки информации и различные информационные системы.

Отвечу – это не то, что я имел ввиду, задавая вопрос: что значит быть Х. Я говорю, что внутреннее качество того, что значит быть Х не может быть раскрыто в отрыве от всего перечисленного Вами.

Например, два независимых аспекта того, что значит быть Х (как это я себе представляю) можно поменять местами без всяких поверхностно открытых целей или причин, без различий в функциональных возможностях видеть, размышлять, принимать решения и т.д. К примеру, понятия «что значит различать цвет неба» и «что значит различать цвет травы» и пр. могут быть взаимно заменены у Х и всё, что угодно ещё, может быть заменено чем-то таким же. Х может сказать нам, что случилось нечто странное, и цвета травы и неба теперь иные, не такие как раньше. Но он не сможет нам сказать, какими они были до этого. Он не сможет определить, когда цветовосприятие у него было такое же, как у других людей сейчас, какое было у них или будет.

Внесём ясность: многие из приведённых здесь высказываний верны, но не все. Различия верного от неверного едва уловимы и их очень трудно выявить. Но нам следует это сделать для обозначения требований к разработке искусственных мыслящих личностей. Одним их таких требованием является то, что человекоподобные, незомбированные роботы должны быть способны к восприятию опрокидываний куба Неккера.

Обращаемые квалиа

Изучим ситуацию, которая происходит при рассматривании куба Неккера: внезапно куб опрокидывается. Хотя изображение на сетчатке и видимая 2D-структура остаются неизменными, 3D-интерпретация уже иная. Линии, а точнее грани куба, которые ранее были направлены вниз от наблюдателя, обратились и уже направляются вверх от него. Верхняя

квадратная плоскость куба, которая ранее простиралась от наблюдателя, теперь простирается к нему.

На примере куба видно, что такого рода обращения можно обнаружить извне и описать для других людей. Вполне возможно, что ученые, изучающие деятельность мозга, выяснят, какие именно неврологические процессы отвечают за этот переход из одного состояния в другое, более того, в будущем будут созданы бесконтактные механизмы идентификации подобного рода переключений.

Также вероятно то, что вскоре органы зрения у роботов достигнут такого совершенства, что смогут улавливать обращения и станет обычным требование к нормальному техническому зрению содержать в себе способность локально управлять двусмысленными фрагментами образа, которые могут иметь различную интерпретацию в зависимости от контекста общей картины. (То, что в одном контексте выглядит как нога, в другом контексте может оказаться рукой).

Что значить для робота обладать способностью к неккеровским обращениям? Это значит (а) ожидать увидеть новый образ в системе линий куба; (б) иметь теоретические объяснения в нейрофизиологии и компьютерной науке; (в) быть способным к внешнему обнаружению (по крайней мере, в принципе); (г) быть объяснённым в терминах геометрических соотношений; (д) иметь совокупность необходимых внутренних и внешних условий.

А может ли обращаться цвет?

Цвет на части поверхности объекта также может обращаться в зависимости от контекста. Это демонстрируют картинки, именуемые «иллюзиями», например, фигуры Каница[8]. Рассматривая их, мы видим замкнутые линии фигур, хотя эти линии не прорисованы. Другим примером могут служить чёрные прямоугольников на белой поверхности – мы видим тёмные области между прямоугольниками, но если приглядеться, темные области осветляются. Имеется ещё ряд примеров подобных явлений. Может, придумают объяснения «иллюзиям», в которых будут обращаться цвета. В них, например, цвет будет «переключаться» с зелёного на красный и, наоборот. Объяснения цветовых обращений будут включать в себя п.п. (а) до (д), характерные для куба Неккера.

Семантические ловушки

Здесь я приведу примеры старых, хорошо известных семантических ловушек, в которые можно попасть [и не выбраться].

Очевидно, (логически) возможно, что эта чашка на столе сдвинется на фут на восток, затем мои очки сдвинутся туда же и тоже на фут, а за тем и стол, и дом, вся Земля, за ней – Солнце вместе со своей солнечной системой, и так далее, пока всё во вселенной не сдвинется на фут на востоку. (Конечно, при таких последовательных перемещениях законы физики временно изменятся, но для наших рассуждений это допустимо, логически допустимо). В конце всего этого ВСЁ во Вселенной окажется смещённым на восток, на фут. Но новое состояние Вселенной будет СОВЕРШЕННО неотличимо от предыдущего.

[8] См., например, http://www.yorku.ca/eye/kanizsa.htm

Можно ли быть уверенным, что такие сдвиги уже не произошли ранее, просто мы забыли об этом в силу ограниченности нашей памяти? Может сдвиги постоянно происходят, т.е. всё потихоньку перемещается на восток, но засечь это невозможно, т.к. все измерительные инструменты и точки отсчёта тоже перемещаются, а законы физики таковы, что никакие эксперименты тут не помогут? (Вспомните эксперимент Майкельсона, его неудачу при попытке определить, куда движется Земля сквозь «космический эфир»).

Если вы и вправду думаете, что имеет смысл то, что я сейчас гипотетически описал и что на самом деле имеются сдвиги во Вселенной, которые невозможно засечь, то можете отложить эту статью в сторону, т.к. вы нуждаетесь в более сильном лечении, чем то, которое я могу вам предложить.

Это – один из примеров заманчивости вымыслов, однако подобные вымыслы сильно укореняются в нашем сознании. Если вам интересно искать ответы на вопросы типа «Действительно ли на Луне наступает полдень, когда она расположена строго над Гринвичем, а в Гринвиче 12 часов дня?», тогда дело совсем плохо, т.к. этот вопрос более не определён, нежели чем вопрос «Какого цвета цифра «9» – жёлтого или зелёного?»

Если вы считаете, что окружающие вас цвета могут вдруг замениться на совершенно другие и что разницу вы не замечаете, то это означает, что такое может происходить постоянно, но вы того не замечаете в силу устройства вашей памяти, которая при каждом подобном изменении забывает то, что было раньше.

Эта идея, в свою очередь, связана с идеей того, что любая пара чувств может быть взаимозаменяема, но обнаружить «замену» не удаётся, так как она осуществляется при полном сохранении всех функциональных аспектов системы.

А в свою очередь последняя идея связана с вопросом – движется ли наша Вселенная непрерывно на восток, запад, северо-восток или ещё куда со скоростью 3 мили/час или 3 см/час, а может с какой-то другой скоростью, но просто это движение незаметно. В (Dennett 1991) рассматриваются схожие проблемы.

Бессвязность знаний это не то же самое, что их нехватка

Конечно, всё вышеприведённое имеет смысл. Эти вопросы стимулируют мышление, рефлексию, интересы и пр., выступающее основой житейского, научного и философского познаний.

Но попытки ответить на эти вопросы не способствуют действительному их решению.

Подобным образом многие не могут разглядеть бессмысленность множества философских вопросов и описаний якобы возможных сценариев их решения.

Похожие вещи лежат в основе религий и многих суеверий.

Выводы

Постановка вопроса «что значит быть X?» и его решение с целью получения детальных ответов, способствует раскрытию внутренней онтологии существ и связано с потребностью компьютерной реализации информационных процессов, лежащих в основе этой онтологии. Над данным вопросом следует задумываться при проектировании и разработке искусственных человекоподобных систем. Его должен обдумывать инженер, создающий новые кибернетические механизмы; программист, разрабатывающий компьютерные игры для детей; учёный пытающийся найти ответы путём проектирования и опытной проверки проектов. Поэтому далеко не всякая критика искусственного интеллекта является необходимой частью [методологии] ИИ, следует учитывать предполагаемые достижения в будущем.

Конечно, многие философы останутся равнодушными к тому, что изложено в этой статье. Они обвинят меня в том, что я закоренелый эмпирик, предсказатель, устаревший позитивист, сциентист, замаскированный зомби и их покровитель, варвар, лжец («дающий нам фальшивую анестезию», как уже сказал один философ), а возможно, обвинять и в чём-то похуже.

Переубедить их невозможно. Когда разум пойман в сетку убеждений, которые верны с точки зрения грамматики и логических выводов и которые действительно что-то значат для этого запутавшегося разума, то никакая обоснованная аргументация не способна изменить ситуацию. Нужна длительная терапия философией. Но и она иногда не помогает.

Я сказал всё это, т.к. знаю, что значит считать, что кто-то понимает то, что сказал ему другой, в то время как третий выдвигает аргументы и задаёт вопросы. Я знаю, что значит быть в этой среде. Я пребываю в ней.

Ссылки

J. Cohen and I. Stewart *The collapse of chaos*, Penguin Books, New York, 1994.

D. C. Dennett, *Consciousness Explained* Penguin Press, Allen Lane, 1991,

J. McCarthy, Making robots conscious of their mental states, *AAAI Spring Symposium on Representing Mental States and Mechanisms* Stanford, 1995, Accessible via http://www-formal.stanford.edu/jmc/

Thomas Nagel What is it like to be a bat, in *The mind's I: Fantasies and Reflections on Self and Soul* Eds D.R. Hofstadter and D.C.Dennett Penguin Books 1981, pp391--403 (Followed by commentary by D.R.Hofstadter, pp403--414.)

Selfe, Lorna *Nadia: a case of extraordinary drawing ability in an autistic child* London, Academic Press, 1977.

Sloman, `On designing a visual system: Towards a Gibsonian computational model of vision' *Journal of Experimental and Theoretical AI* 1,4, 289-337 1989
(Also available as Cognitive Science Research paper 146, University of Sussex).

Sloman, `The mind as a control system', in *Philosophy and the Cognitive Sciences,* (eds) C. Hookway and D. Peterson, Cambridge University Press, pp 69-110, 1993 (Supplement to *Philosophy*)

Sloman, (1994) Semantics in an intelligent control system, in *Proc. British Academy and Royal Society Conference: Artificial Intelligence and The Mind: New Breakthroughs Or Dead Ends?* in *Philosophical Transactions of the Royal Society: Physical Sciences and Engineering,* Vol 349, 1689, pp 43-58 1994

Sloman, (1995) Musings on the roles of logical and non-logical representations in intelligence, in Janice Glasgow, Hari Narayanan, Chandrasekaran, (eds), *Diagrammatic Reasoning: Computational and Cognitive Perspectives*, MIT Press, 7--33 *Фланаган*

Зомби и роль сознания

Оуэн Фланаган, Томас Полджер[1]

Перевод Татьяны Кураевой

Аннотация

Мысленный эксперимент Тодда Моуди – «Земля зомби» – представляет неудачную попытку опровергнуть принцип «несущественности сознания». Эксперимент следует пересмотреть. Мы защищаем от критики принцип «несущественности сознания» и утверждаем, что мысленный эксперимент с зомби выдвигает на первый план самые сложные проблемы в изучении сознания - необходимость объяснения самого феномена сознания и какую(ие) функцию(и) сознание выполняет.

1. Принцип несущественности сознания

Принцип *несущественности сознания* (*сознательного инэссенциализма*) – это точка зрения, согласно которой, хотя в любой области знаний d любая интеллектуальная деятельность i выполняется мною сознательно (при условии сопровождения i сознанием), в принципе, я могу осуществить её бессознательно (без этого сознательного сопровождения). Среди всего прочего, *принцип несущественности сознания* усиливает вопросы о роли и, следовательно, об адаптивной и эволюционной значимости сознания, так как заявляет, что сознание не представляется необходимым ни в метафизическом, ни в логическом отношении. Оно не необходимо в возможных мирах, где обитают функционально эквивалентные нам, [людям], существа. Однако в нашем, реальном мире, сознание – суть реальность. Почему? Какое адаптивное преимущество получает существо, которому даровано сознание в отличие от бессознательного существа, которое, кстати, может быть чрезвычайно интеллектуальным. Может, сознание ничего и не даёт человеку?

Принцип несущественности сознания намного правдоподобнее концепции философского зомби – несчастного болвана, часто фигурирующего в т.н. «войнах сознаний», в которых он предстаёт поведенчески неотличимым от нас, людей, несмотря на полнейшее отсутствие у него феноменального опыта. Зомби – это просто автомат. Быть зомби - «значит не быть». Поведение зомби подобно нашему, разумному поведению, однако оно полностью неразумно в плане сознания. Тем не менее, зомби способны обмануть самые совершенные «ментальные детекторы».

[1] Оуэн Фланаган (факультеты философии, психологии, и нейробиологии, Университет Duke), Томас Полджер (факультет философии, Университет Duke), 1995 г

Проблема зомби преследует несколько различных целей. Она в [предельной степени] поднимает вопрос о роли сознания. Она ярко демонстрирует традиционную проблему других сознаний, которая заключается в следующем: как можно быть уверенным в том, что некоторые, а может быть, и все окружающие нас люди – не зомби? Проблема зомби также часто используются для раскрытия несостоятельности функционализма, и косвенно, для опровержения теста Тьюринга.

Согласно защитникам ортодоксального теста Тьюринга, если машина и человек воспроизводят те же самые отношения ввода – вывода, то есть если человек и машина функционально тождественны, тогда, приписывая интеллект человеку, мы обязаны приписать интеллект и машине.

Согласно критикам функционализма, ортодоксальный тест Тьюринга можно использовать для проверки интеллектуального поведения, но отнюдь не для проверки на наличие сознания. Проблема зомби это и раскрывает.

Некоторые защитники теста Тьюринга пытаются исправить это возражение в поиске более сильного испытания, нежели тест Тьюринга. Они предлагают Тест Тьюринга «с мозгом» («Brainy-Turing-Test»). В дополнение к поведенческой эквивалентности теперь требуется эквивалентность нервной системы: то есть помимо тождественности ввода-вывода нужна тождественность с мозговой активностью.

Неприятно то, что даже при данном, более строгом испытании, непоколебимым остаётся положение о мыслимости (conceivability) зомби. Данное положение заключается в том, что можно помыслить существ, у которых и поведение и мозг эквивалентны нашим, однако эти существа сознанием не обладают. Например, направляя ячейки фоторецептора на то, что мы называем красным пятном и говорим при этом «красный», мозг такого существа может пребывать в том же состоянии, в каком пребывает человеческий мозг в состоянии «наблюдение красного», однако у зомби будет отсутствовать опыт наблюдения красного — у него вообще нет никакого опыта, несмотря на то, что [феноменальный] опыт – как это утверждается сторонниками теста Тьюринга «с мозгом» – соотносится с состояниями мозга. Во-вторых, требование тождественности нервной системы и [сознательного] опыта представляется шовинистичным. Искусственные сердца и почки возможны [субстратно отличные от человеческих – Т.К], почему же невозможен искусственный мозг (или нечто биологически отличное от [человеческого мозга]), осуществляющий то же самое, что выполняет и подлинный мозг?

Первый пункт показывает, что добавление требования эквивалентности нервной системы недостаточно для получения эквивалентности опыта. Второй случай показывает, что в этом даже нет необходимости. Так что тест Тьюринга «с мозгом» не справляется с задачей идентификации сознания [у системы] – наличия или отсутствия у неё сознания.

2. Зомби Моуди

В недавней работе «Беседы с Зомби» Тодд Моуди выдвигает аргументы против правомочности принципа несущественности сознания хотя бы потому, что «зомби, которые пребывают среди людей, способны сносно овладеть нашим языком, включая философские диалоги о сознании [и других ментальных понятиях]. Но мир зомби, [в котором нет людей] породить такие понятия не может» [С.199].

Последовательно, шаг за шагом, разберем эти аргументы. Шаг первый: следует признать возможность тьюринг-идентичных зомби, растущих в нашей среде. Например, можно вообразить фантастически совершенную робототехническую версию NET-talk –коннекционистской машины, которая обучается языку в окружении людей и затем использует его так же, как мы его используем, включая термины «верю», «надеюсь», «вижу» и т.д.

Второй шаг отрицает, что некоторый зомби или изолированное население зомби *могут* «породить» наш менталистский словарь (словарь ментальных терминов). Следует учесть многозначность слова «мочь», так что следует быть внимательными. Является ли утверждение о таком словаре метафизически/логически невозможным, отрицая при этом, например, возможность происхождения слова «сознание» и производных от него слов среди существ, которые не осознают [смысла] этого слова. Или это утверждение является номинально (nomically) невозможным в том плане, что появление такого словаря нарушает закономерности нашего мира. Или это утверждение просто «невероятно»? С той же степенью вероятности ментальный словарь появится среди изолированной группы зомби, с какой папа Римский и мать Тереза в следующем месяце проведут пикник на Луне? А пикник на Луне и метафизически и логически возможен – Н. Армстронг и НАСА доказали, что законы природы не противоречат тому, что люди могут побывать на Луне.

Таким образом, слово «возможность» многозначно. По всей видимости, никто под «возможностью» не принимает возможность реального существования зомби в нашем мире – т.е. того, что зомби живут среди нас, хорошо маскируясь при этом. Возможность существования зомби, поведенчески неотличимых от людей, понимается в метафизическом, логическом и номинальном смыслах. Все эти разновидности понятия «возможности» рассматриваются из-за того, что они уточняют

теории о природе, функции и критериях сознания. Если системы «такие же, как мы» могут существовать без сознания, то зачем тогда вообще нужен этот придаток [, т.е. сознание]? Производит ли сознание нечто такое, что не может производится без его помощи? И, кроме того, если сознание всё-таки востребовано, воздействует ли оно на вселенную?

3. Земля Зомби

Для определения того, как Моуди понимает «возможность», утверждая, что принцип несущественности сознания требует пересмотра (199), следует более внимательно рассмотреть детали его мысленного эксперимента. Моуди предлагает вообразить планету, населённую одними только зомби:

Предположим мир, в точности похожий на наш, помимо одной детали: люди этого мира бессознательны. У них, как и у нас, сложное поведение, включая речь. Однако у этих людей поведение не сопровождается никаким сознательным опытом. (196)

Данное положение, [считает Моуди] удобно обсуждать, мысленно представив, что на Земле Зомби каждый из нас имеет зомби-двойника.

Моуди утверждает, что люди и жители Земли Зомби будут поведенчески различаться. Различия проявятся, в первую очередь, на «уровне речевой коммуникации» (197). У зомби будут отсутствовать понятия «мечтать», «болеть», «видеть» и т.п., так как у них отсутствует внутренняя [субъективная] жизнь, являющаяся референтом наших ментальных терминов. Отсутствие таких понятий, утверждает Моуди, проявится в языке Земли Зомби. Слова, обозначающие сознательные явления никогда не будут изобретены. Жители Земли Зомби не будут пользоваться ментальными терминами. Отсутствие в языке таких терминов и станет основным отличием их от людей, что будет служить критерием «зомбиевого» (zombiehood), [в отличие от «человеческого»] (199).

Моуди предполагает, что если принцип *несущественности сознания* верен, тогда возможно не только существование мира зомби, подобного нашему (за исключением, конечно, того, что он населён зомби [, а не людьми]), но и возможно развитие такого мира до полного соответствия нашему, но в котором будет отсутствовать сознательная жизнь. Метафизика и логика допускают подобного рода возможность. Кроме того, [естественные науки,] законы природы, также не накладывают особых запретов.

Поэтому второй шаг влечет не просто возможность того, что можно вообразить факт существования Земли Зомби и представить себе коннекционистских зомби (это оказалось допустимым ещё на первом шаге), растущих в окружении людей, но и возможность вообразить целый мир зомби, развивающийся согласно законам эволюции.

Однако следующий, третий шаг приводит к провалу этого мысленного эксперимента. Необнаруживаемые, однако, эволюционирующие зомби невозможны. Они невообразимы в силу того, что жители Земли Зомби должны отличаться от людей, т.к. по причине отсутствия сознательной жизни зомби никогда не будут использовать и, более того, никогда не смогут создавать ментальные понятия и соответственно, словарь [ментальных терминов], подобный тому словарю, который имеется у нас. Согласно Моуди, даже при всём при том, что «речевая активность о» во время, например, сна или в процессе зрения «не требует сознательности, эти понятия должны появляться в языковом сообществе» (199). Таким образом, «критерий зомбиевости» на Земле Зомби – это отсутствие какого-либо психологического словаря.

4. Как вообразить Землю Зомби?

Допустим, Землю Зомби можно вообразить. К чему это приведёт? Моуди считает, что «если принцип несущественности сознания верен, тогда мы не смогли бы определить, являются ли, например, пришельцы из другого мира зомби или нет. В конце концов, если нет никакого необходимого поведенческого различия между ними и нами, как того требует данный принцип, то тогда не будет никакого опознавательного «клейма зомби». (197)

Данный аргумент можно перефразировать так:

Если принцип несущественности сознания верен, тогда нет возможности отличить зомби от сознательного существа.
Однако имеется возможность утверждать, что жители Земли Зомби – это зомби.
Следовательно, принцип несущественности сознания неверен или требует уточнения.

Аргумент формально корректен и явно выражен в форме [логического] правила modus tollens. Однако он не обоснован. Первая предпосылка ложна. Сознательный инэссенциализм утверждает лишь то, что для любого интеллектуального акта i, выполненного в любой познавательной области d, даже если мы [люди] и делаем i с осознаванием этого акта, то акт i, в принципе, может быть осуществлён без всякого его осознавания. Тезис ничего не говорит о ситуациях, в которых присутствие или отсутствие сознания на самом деле можно или нельзя будет обнаружить.

Вторая посылка - утверждение, что зомби на Земле Зомби можно легко обнаружить. Если зомби на самом деле такие, как их обрисовал Моуди – то есть у них нет ментальных терминов, то тогда посылка верна в плане возможности. Однако эта посылка подкреплена модальной

логикой, то есть она или возможна или необходима. Если Моуди считает, что обнаружить зомби на Земле Зомби принципиально невозможно — в противоположность тому, что это неправдоподобно (то есть зомби просто необнаруживаемы), тогда Моуди неправ и вторая посылка также ложна.

Это легко представить. Жители Земли Зомби сообразительны и «информационно чувствительны», однако, будучи зомби, - не «осознают опыт» (Flanagan 1992). Зомби издают речевые сигналы и фиксируют закономерности своего мира посредством языка, схожего с нашим. Согласно Моуди, это означает, что их математика и естествознание «вероятно, будут очень похожими на нашу математику и естествознание» (197).

Однако зомби не будут создавать и применять речевые сигналы и письменные символы для закономерностей, которых нет в их мире. Они не будут обозначать ментальные образы сознания [, которого нет]. Земля Зомби лишена этих образов и так как зомби не нуждаются в терминах для обозначения того, что не существует, «ментальные слова» никогда не появятся на Земле Зомби.

Что и говорить, в крайней степени маловероятно и чрезвычайно неправдоподобно появление менталистского словаря среди зомби Моуди. Однако является ли этот словарь метафизически, логически и номинально невозможным? Нет. По крайней мере, есть два пути, по которым зомби могли бы прийти к использованию ментальных терминов. Первый связан с культурным развитием зомби. Второй – с развитием зомби как рода.

По поводу культурного развития может показаться очевидным, что зомби, не являясь субъектами опыта, не будут иметь никаких причин создавать термины, которые мы могли бы перевести словами, например, «сон» и «зрение». Но рассмотрим следующую возможность. Зомби, снабжённые информационными датчиками, фиксируют (но не воспринимают) деревья на Земле Зомби. Они фиксируют тот факт, что зомби-соотечественники иногда сталкиваются с деревьями. Фиксация закономерности столкновений проводит к генерации предупреждающего сигнала типа «Осторожно!». То есть естественная социальная функция заставляет зомби поворачиваться так, чтобы их фоторецепторы воспринимали релевантную информацию об изменении направления движения. Зомби присваивают термин направленности фоторецепторов в правильном направлении «зрением». Безусловно, это – зрение[3], неосознаваемая копия с нашего зрения[2]. Дело в том, что мы не знаем, что это есть *зрение*[3], а не *зрение*. Также мы не знаем и того, что именно зомби могли бы обозначить термином *зрение*.

[2] Здесь и далее ментальный термин зомби помечен индексом «з» – Прим.перев.

Можно придумать такую же историю для снов. Зомби перезаряжаются во время того, чему они присваивают термин «сон». Ночами, бывает, происходят сбои в процессе перезарядки. Предположим, эти неисправности приводят к активации речевых центров зомби.

Как в случае с деревьями, флоры и фауны вообще, и движений среди них соотечественников, у зомби можно обнаружить релевантные закономерности. Утром они отчитываются о последовательности операций по активации центра речи (хотя это информирование лишено для них какого-то смысла): «Я видел дерево. Джейн была на нем. На ветвях мы занимались любовью». Такие отчёты зомби начинают обозначать термином «сны». Конечно, это - *сны*^{з}, а не *сны*, но дело в том, что такие термины, вполне могли появиться у зомби на Земле Зомби. Отсутствие таких [ментальных] терминов или их эквивалентов, способных к переводу в рамках лингвистической коммуникации зомби не может служить в роли «клейма зомби». Ведь эти термины могли же появиться, в конце концов!

Отчеты о снах^{з} могут сослужить зомби хорошую службу. Данные, активизированные во время сна в речевом центре, могут стать схемой дальнейшего поведения зомби или информировать о качестве перезарядки. Активация речевых центров может стать подспорьем и препятствием для следующей перезарядки – и это зомби будут узнавать^{z} и артикулировать.

По отношению к видовой эволюции зомби, Моуди допускает мысль, по крайней мере, в рамках своего мысленного эксперимента, что [законы] эволюции не отрицают возможность существования Земли зомби. Здесь он, кажется, подразумевает все виды возможности - не только метафизическую и логическую, но и номинальную возможность. Законы природы, насколько нам известно, допускают возможность развития интеллектуальных, информационно-восприимчивых, но несознательных существ.

Нео-дарвинистская теория подразумевает не только приспособление к условиям существования, но и свободу действия организма, а также счастливый случай. Нет никакой метафизической, логической или номинальной невозможности в идее того, что мать-природа вдруг взяла и случайно создала для зомби врожденный язык мышления с лексическим пространством поведенческих дескрипций упомянутого выше вида. Этот язык восполняет ряд физических недостатков (термин «зрение» означает «поворот фоторецепторов в требуемом направлении»; «сон» - это «активизации разговорного центра во время сна»). Может показаться, что эти [восполняющие] термины имеют ментальные референты. На самом деле это не так, таких референтов не существует.

5. Философия Зомби

Сделаем небольшую уступку и допустим, что зомби на Земле Зомби всё же создали менталистский словарь. Здесь Моуди приготовил ещё одно опровержение. Допустим, - утверждает он, - можно представить, что зомби произносят такие предложения, как «я вижу красный», и «я спал, во сне ко мне явилась кузина Джейн и мы на дереве занимались сексом». Однако невозможно вообразить зомби, философствующих так же, как мы философствуем о снах и зрении. Зомби не способны создать и философию разума. Неспособность философствовать и будет являться отличительным признаком зомби, «клеймом зомби»!

Описывая сценарий развития Земли зомби, Моуди размышляет о философии зомби. Он заявляет, что у зомби не могу возникнуть ни ментальные термины, ни философия разума. По первому пункту он оказался неправ. Что относительно второго? Снова неверно.

Моуди предполагает, что философы-зомби «особенно могут быть увлечены^{z} человеческой философской литературой о снах». (198) Но они не будут увлекаться^{z} фантазиями и грёзами вообще – ведь их сны^{з}, как это мы представили выше, ограничены утренними отчетами об активизации речевого центра [после перезарядки]. Что и как говорят зомби о снах^{з} совершенно отличается от того, что и как мы говорим о снах. Будут ли Зомби мечтать? Нет. Но как мы предположили, они будут использовать слово «мечта», чтобы говорить о своих мечтах^{з} и задавать вопросы по поводу мечты^{з}, подобные нашим вопросам о мечте. Им интересно^{з}, к примеру, будет знать следующее: то, что они называют «грёзами»^{з} – это отчеты речевого центра во время сна или в период пробуждения. Философы-зомби зафиксируют различия в работе речевого центра при пробуждении и в ходе обычного поведения. На основе этого они различат соответствующие образы. Например, многие схемы поведения, зафиксированные во время сна и представленные в отчетах речевого центра не будут совпадать со схемами поведения, которыми «нормальный зомби» руководствовался бы в свете дня.

Подобным образом разрешается проблема инвертированного спектра, которая для зомби была бы проблемой суждения об инвертированном цвете. И для нас и для зомби эти проблемы [, в принципе] тождественны. Предположим, нормальные зомби, после наблюдения световой волны с длиной x, переходят в состояние диспозиции к выдаче фразы «объект зелёный» и далее они действуют в соответствии с этой диспозицией. Все, что требуется для решения проблемы суждения об инвертированном спектре состоит в двойном изменении поведенческих схем. Для нас (то есть для случая традиционной проблемы инвертированного спектра) одна схема предполагает качественное [осмысленное] восприятие цвета, другая схема – акт речи. У зомби решение

достигается на уровне бессознательных инверсий — никакое «внутреннее восприятие» (198) для этого не требуется. Сначала, при наблюдении объекта, испускающего волну длины *x*, зомби, выносящий суждение о перевёрнутом спектре, переходит в диспозицию, в которой нормальный зомби говорит: «[вижу] красный объект». Если при этом поменять соответствующие провода, зомби переходит в диспозицию, в которой выносится суждение «вижу зелёный объект». Таким образом, двойная инверсия проводов решает проблему, тождественную проблеме инвертированного спектра.

Ранее мы упомянули, что мысленный эксперимент с зомби используется для демонстрации проблемы других разумов. Зададимся вопросом – а могли бы сами зомби поднять тот же философский вопрос^{Z} ? Да, зомби могли изучать[3] проблему, которую они могли бы назвать, достаточно странно для себя - «проблему других разумов». Как это возможно?

Вспомним, что единственное различие между Землей зомби и Землей [людей] – состоит в том, что её жители - зомби. Важно то, что нет никакого повода думать, что зомби убеждены[3] (не говоря о том, что они полагают[3]), что они есть зомби в нашем, [человеческом] смысле этого слова. На самом деле, в соответствии с нашей гипотезой, они могли бы самих себя назвать «сознательными» [существами]. Даже после встречи с нами зомби так и не поняли бы, что они зомби. Они были бы точно в таком же положении относительно нас (и друг друга), в каком мы находимся относительно них (и друг друга): они не имели бы никакого подлинного поведенческого доказательства, что мы чем-то отличаемся от них. Но всё это отнюдь не по причине того, что зомби не могут различить «сознательных, себе подобных существ» от «зомби».

Допустим, один философ-зомби, размышляющий^{Z} о грёзах и цветовой инверсии, подходит к постановке более основательного вопроса. Философ-зомби может полагать то, что он «мечтает» и «видит» объекты. Он может также знать кое-что о мозговых состояниях, которые вызывают его «мечты» и «видения». И он фактически может подтвердить, что такие зависимости имеют место тогда, когда он действительно мечтает^{Z} и умозрит^{Z3}. Способность грезить и способность к умозрению, частично, и есть то, что он имеет в виду, называя себя «сознательным».

Философ-зомби мог бы задать вопрос – являются ли все другие зомби сознательными точно такими же, как и он, сознательным. Реальны ли мечты других настолько, насколько реальны отчеты его центра речевой активности (и позвольте, со стороны центра визуальной активности), отчеты, в которых фиксируются состояния во время быстрого [, утреннего] сна. Либо так называемые «мечты» других зомби – суть нечто

3 Sight – «видеть образы», переведено на типично русский манер – «умозреть» - Прим.перев.

сверхъестественное подобно ясному состоянию ума сразу после пробуждения. Как только [в памяти] зомби сформируется шаблон для сравнения, то он легко может выразить проблему *других зомби* [курсив мой – Т.К.] в таком виде: наши соотечественники «осознают» себя подобно мне или же все они – «зомби»? Философ-зомби задаст вопрос – не являются ли его соотечественники «зомби», у которых мозговые состояния продуцируют «правильное» поведение и которые совершенно не осознают того, что делают. Таким образом, имеется достаточно доводов для демонстрации возможности возникновения у зомби «проблемы других разумов».

6. Трудная Проблема

Мы только что защитили принцип несущественности сознания от мысленного эксперимента Земли Зомби. Тезис данного принципа - заявление о возможности осуществления всякой интеллектуальной деятельности без осознавания этой деятельности – был рассмотрен при различных толкованиях слова «возможность». Если была бы истинной логическая невозможность мира существ, которые функционально неотличимы от нас, являясь при этом бессознательными, то тогда *принцип несущественности сознания* был бы ложным или требовал бы уточнения.

Однако, согласно вышеизложенным доводам, ни метафизика, ни логика, ни даже законы природы не препятствуют реализации принципа несущественности сознания [зомби]. Зомби побеждены.

Настало время спуститься на нашу землю с Земли зомби и с разреженных высот модальной логики и мысленности (conceivability) мысленных экспериментов. Пришло время напомнить, зачем философам проблема зомби и почему расплодилось так много зомби в ходе всякого рода мысленных экспериментов. Как мы сказали в начале, есть несколько целей, которые преследуют мысленные эксперименты с зомби: такие эксперименты либо подтверждают либо опровергают заявления функционалистов, защитников и противников теста Тьюринга, искусственного интеллекта, и тех, кого волнуют перспективы искусственной жизни. Размышление о возможности зомби проливает свет на проблемы философии языка, особенно, по поводу непостижимости сути референции или, например, на проблемы онтологической релятивности.

В заключение предлагаю сосредоточиться на проблеме, трудность которой недооценена, но к которой Земля Зомби привлекает особое внимание. Это – вопрос «почему мы обладаем сознанием?».

Первоначальное обсуждение принципа несущественности сознания идёт вразрез со следующим вопросом: «Сознание не должно возникать. Можно помыслить то, что эволюция приведёт к созданию существ, столь же эффективно действующих и интеллектуальных, как и мы, а может, даже более эффективных и интеллектуальных. Однако эти существа не владеют

предметами опыта... Из этого факта – того, что сознание несущественно для высокоразвитой интеллектуальной жизни, - следует, что сознание несущественно и для нашего, специфически человеческого типа интеллектуальной жизни « [6]

Предположим, с одной стороны, что Homo Sapiens сознательны. Это верно. Но с другой стороны верно и то, что нет никакой метафизической, логической или номинальной необходимости в создании нас такими. Почему мать-природа решила, что обладать опытом – это хорошая стратегия для нас, людей, и, возможно, для других многочисленных млекопитающих и животных?

Некоторые философы, например, Давид Чалмерс, Джозеф Левин, Колин МакГинн, Томас Нагель, считают, что самая трудная проблема состоит в объяснении причинной зависимости между мозговыми и феноменальными состояниями. Это - действительно трудная проблема. Но она, конечно, не более трудна, чем проблема того, почему и как происходит, что существа вообще начинают что-то осознавать. Почему венцом эволюции стали существа, у которых имеется нечто большее, помимо того, что имеется у просто информационно-восприимчивых существ? [Сегодня] отсутствуют приемлемые теории, решающие данную проблему, что служит причиной беспокойства по поводу принципа несущественности сознания.

Конечно, читатель может посчитать, что мы шутим. Ведь должны же быть нормальные теории, объясняющие возникновение сознания. Можно сколь угодно долго и где угодно искать такие теории – не найдёте. Почему мы так считаем, мы не можем ответить путем анализа всех таких теорий. [Их слишком много]. Рассмотрим лишь некоторые теории наших недавних оппонентов: в первой дается объяснение явлению сознания вообще с позиции адаптации. В двух других объясняются специфические формы сознания – осознание боли и чувство вожделения [страсти, похоти]. Последняя теория обращается к сознанию как к объяснению некоторых сложных ментальных способностей.

Поток сознания. Часто можно слышать, что система «разум/мозг» - это «параллельная распределённая система», способная одновременно осуществлять много различных вычислений. Представление о сознании как последовательном потоке «одно-одновременно» - выглядит блестящей стратегией предоставлять «центру управления всей системой» только ту информацию, которая системе на самом деле нужна. Систему окружает много всего того, о чём центру управления знать не обязательно, зачем его зря беспокоить [мелочами], ему нужна только самая важная информация самого высокого качества. Данная идея и интуитивно и теоретически привлекательна. Ведь мы знаем, что в каждый момент внутри нас очень много всего происходит, что полностью мы не осознаем. И кажется, что наш поток сознания, мощный и стремительный предназначен для наведения порядка в хаосе параллельной обработки информации.

Допустим, что данное предположение – хороший ответ на проблемы, подобные проблеме узких вычислительных мест в последовательном нейродинамическом процессе и проблеме информационной перегрузки. Однако это предложение ничего не объясняет по поводу того, почему центром управления (ко всему прочему, гипотетическим) востребована именно сознательная информация, а не беспристрастная информация типа той, которую обрабатывают ничего не сознающие PDP-компьютеры.

Боль. Деннет (1990) пишет: «вне всякого сомнения, имеющаяся в мозгу сигнальная система болевых волокон и взаимодействующих трактов является эволюционным приобретением, даже если за него следует платить цену болью, с которой мы не можем ничего поделать». И далее задаёт прямой вопрос: «Но почему боль должна ранить столь сильно? Почему, например, она не может быть просто громким звонком в ухе разума? « (1990, 61).

Деннету, как никому другому, крайне близка проблема адаптивной роли сознания. Но как и все другие, он не предлагает удовлетворительного ответа на поставленный вопрос или на некоторую его разновидность, например: почему система не может быть устроена таким образом, чтобы определять нежелательное для себя посредством некоторых врожденных средств, таких как (среди других прочих) температурные датчики и сенсорные датчики уровня крови. Эти средства, в свою очередь, могут быть подключены к эффекторным средствам реализации соответствующих действий.

Вожделение. Тот же самый аргумент подходит и для более приятных опытов. Сразу после вышеприведённых размышлений о боли Деннетт пишет: «Полезны гнев, страх и ненависть (и я настаиваю, что эволюционная польза вожделения не нуждается ни в какой защите)». Почему для Деннетта очевидна эволюционная польза вожделения? Вероятно, он рассуждает так: учитывая, что мать-природа заботится только о генетической пригодности для воспроизводства [вида] (и которое не все виды вовлекает в сексуальное воспроизводство - оставим любовные утехи), то хорошей идеей было бы создание мощного стимула, в качестве которого могло бы быть желание секса только с пригодными для воспроизводства партнёрами. Вожделение [- это и есть такое желание] задаёт мощный стимул и усиливает потребность в воспроизводстве [вида].

Многое кажется правильным. За исключением малой неприятности – перед одним из очередных «очевидных пунктов» о роли вожделения Деннет пишет: «нет почти ничего сексуального (в человеческом понимании сексуальности) в половой жизни цветов, устриц и других простых форм жизни». (Деннетт, 1990, 173). И далее указывает, что самосознание, и более того, вожделение - «среди самых глубоких проблем в современной эволюционной теории» (1990, 172).

Мы должны воспроизводиться. Это верно. Но почему так непосредственно очевидна эволюционная польза вожделения? Обоняние, например, также часто фигурирует в [описаниях] жизни животных, которые только этим и отличаются в плане сексуальной озабоченности. Мужские особи ночных бабочек чувствительны к феромонам, которые выделяют женские особи. Обоняние позволяет им лететь правильно целые мили к местоположению [особи] для спаривания. (Люди, как известно, путешествуют дальше и с меньшей поддержкой). Тогда может главную роль в воспроизводстве следует присудить не вожделению, а обонянию, если учесть факт регуляции спаривания ночной бабочки посредством обонятельных сигналов. Распространяющиеся по воздуху ароматы фиксируются датчиками химического состава. По этим датчикам определяется время спаривания и срабатывает похоть. Может, так? Сегодня у нас нет никакого вразумительного ответа на подобный вопрос.

Обучение и адаптация. Можно счесть возможным, к примеру, и то, что боль и секс регулируются не бессознательными информационными процессами, а потребностью организма в обучении. Если зомби среди нас, то почему бы некоторым вещам не распространять феромоны, которые будут отвращать зомби от ненужного спаривания? Или, например, вместо того, чтобы испытывать чувство боли, какой-нибудь зомби для спасения своего рода касается огня, определяет высокую температуру и выдаёт сигнал «Беги и спасайся!»?

Иногда необходимость сознания рассматривается с позиции адаптации организма к внешней среде. Конечно, секс «по проводам» или детекторы вреда [от пожара] могут оказаться полезными для запуска процедуры-секс-инициации или процедуры-вреда-остерегания. Однако такое «чувство сексуальности» или «чувство боли», по сути, основанное на врожденных инстинктах представляется опасным. Врожденные инстинкты на эволюционно-полезные или эволюционно-опасные события, конечно, заставляют спариваться с кем надо или избегать угроз. Однако они не позволят нам узнать о случайных особенностях нашей окружающей среды, ведь окончательный успех основан именно на возможности преодолевать случайности. [А для этого требуется сознание].

Данный довод не срабатывает. Приспособление, обучение и т.п. для нас не интересны. Они могут осуществляться бессознательно. Например, компьютеры демонстрируют, что обучение и адаптация, как это делают люди, понимания не требуют.

Можно предположить, что сознание людей предназначено не для обучения как такового, а для облегчения обучению. Такая функция сознания радикальным образом изменяет поведение, так как действует на самом высоком уровне – уровне мыслительной деятельности.

Но всё это – частные теории о роли сознания. Изучение проблемы существенности сознания – нужно или не нужно нам сознание – должно

быть освещено в свете вопросов: почему сознание наличествует, почему оно возникает, [есть ли зомби?], участвовали ли зомби-гоминиды в борьбе за выживание и потерпели ли они в ней поражение, а если не потерпели и до сих пор существуют среди нас, то почему выжили?

Крайне трудно придумать убедительную историю об адаптивной роли сознания. Результатом любой истории будет отсутствие объяснений - почему сознание появилось в нашем мире. Хотя мы и пытались рассказать историю, но пока не придумали, почему появились существа, воспринимающие явления опыта, почему они победили или должны победить в эволюционной борьбе у высоко-интеллектуальных, информационно-восприимчивых зомби-подобных организмов. По крайней мере то, что не сделано на пути построения убедительной теории – мы так и не получили ответа на вопросы: почему сознательные субъекты утверждаются в мире и почему мы - не зомби.

Отсутствие такой теории - одна из причин, почему мы нуждаемся в проблеме зомби. Зомби раскрывают много интересного в философии разума. Например, может мы – зомби, а может быть – и нет. Но печально [не то, что мы – не зомби, а] то, что не можем объяснить - почему мы не зомби.

Данная проблема - проблема возникновения и развития сознания – прерогатива биологов, ученых в области когнитивных наук и философов, специализирующихся в философии разума. Проблема требует экспертизы их теорий в рамках эволюционной парадигмы. Требуется междисциплинарное сотрудничество также и других исследователей - психологов, этнологов, палеонтологов, зоологов, и неврологов. Трудно представить, какой вклад в последующем внесут исследователи искусственного интеллекта[4]. Они уже сейчас помогли, поставив на повестку дня вопрос о создании зомби в нашей среде. Они также показали нам, что проблемы интеллекта и сознания сводятся к решению двух трудных проблемы: о причинной зависимости сознания от состояний материи и проблемы возникновения и эволюции сознания. А это требует, для начала, детального изучения сознательных систем - биологических существ, эволюция которых привела к тому, что они стали субъектами сознательного опыта. Требует изучения нас с вами.

---

4 Авторы употребили не термин «artificial intellegence» (искусственный интеллект), а термин «artificial intellegentsia» (искусственная интеллегенция). Это соответствует духу статьи про зомби. Интеллегенция обозначает, исходя из древнегреческой традиции некий самосущий разум, интеллектуальное существо как таковое, включая и мыслящий себя ум. Искусственный интеллект – это приписываемая машине интеллектуальная способность. Зомби лучше соотносить с первым случаем словоупотребления – прим.перев.

СОДЕРЖАНИЕ

Научное издание

ТЕСТ ТЬЮРИНГА. РОБОТЫ. ЗОМБИ.
Сборник переводов с англ.

Под редакцией кандидата философских наук
Алексеева Андрея Юрьевича

Оригинал–макет: Андрей Сочилин
Художник: Василий Алексеев
Технический редактор: Татьяна Кураева
[redacted]: Филипп Петров

Подписано в печать с оригинал-макета 23.03.2006 г.
Формат 60x84/16. Ризография. Гарнитура Таймс.
Усл.печ.л. 7,5. Уч.-изд.л. 12,25. Тираж 100 экз.
Заказ № 80 от 04.04.2006 г._____

Издательство «ИИнтеЛЛ»
109518, г. Москва, ул. Грайвороновская, д.20 - 171
book@aintell.ru
http://aintell.ru

Московский государственный институт электроники и математики.
109028, г. Москва, Б. Трёхсвятительский переулок, 3/12.
Отдел оперативной полиграфии Московского государственного института электроники и математики.
113054, Москва, ул. М. Пионерская, 12