ФИЛОСОФСКИЕ ПРЕДПОЛСЫЛКИ ТЕСТА ТЬЮРИНГА ДЛЯ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА.[1]

Ладов В.А.

В этой статье обсуждены принципы теста Тьюринга для искусственных интеллектуальных систем. Известно, что А.Тьюринг намеревался избежать философских дебатов о сущности построения его машин. Автор этой статьи пытается продемонстрировать, что в любом случае нельзя проигнорировать философские вопросы, связанные с выработанными А.Тьюрингом принципами тестирования систем искусственного интеллекта.

THE PHILOSOPHICAL PRESUPPOSITIONS OF TURING'S TEST FOR ARTIFICIAL INTELLIGENCE SYSTEMS.

Ladov V.A.

The principles of Turing's test for artificial intellectual systems are discussed in this article. It is known that A. Turing intended to avoid philosophical debates about essence of the reason by means of the specific test. The author of this article tries to demonstrate that one cannot ignore philosophical questions in the process of formulating of the test in any case because some interpretations of the reason's essence are the presuppositions of 'Turing's interval' that is the basis of the test.

Данное исследование посвящено обсуждению вопроса о принципах осуществления широко известного в области искусственного интеллекта теста Тьюринга, на основании которого делается заключение о том, является ли та или иная система интеллектуальной. Известно, что в своих намерениях создать такой тест Тьюринг был мотивирован желанием прекратить философские дебаты о сущности разума и установить простой бихевиористский критерий интеллектуальности. В связи с этим мы попытаемся показать, что обойти философские вопросы при формулировке теста все равно не удается, ибо определенные интерпретации

[1] Статья написана при поддержке РФФИ. Грант № 06-06-80003а

сущности разума лежат в основе того «интервала Тьюринга», по которому выстраивается сам тест.

В своей широко известной статье «Computing machinery and intelligence» [1] А. Тьюринг сформулировал настолько простой и внятный критерий определения разумности системы, который, как казалось, должен положить конец всем метафизическим спекуляциям на эту тему. Тьюринг предложил сыграть в игру-имитацию. Пусть испытуемый получает письменные ответы на свои вопросы от двух респондентов, один из которых человек, а другой – машина (компьютер, система искусственного интеллекта). Если испытуемый в прочитываемых ответах не может отличить высказывание человека от высказывания искусственно созданной технической системы, то будем считать такую систему разумной.

При формулировке своего теста Тьюринг оригинальным образом применил уже достаточно распространенную в психологии бихевиористскую тактику. Бихевиоризм получил свое развитие в психологии на фоне усталости и явной слабости современной науки в вопросах фиксации сущностных свойств сознания, его происхождения на основе нейрофизиологических процессов в головном мозге, принципов его функционирования. Оставив в стороне эти сложные и неразрешимые на настоящем этапе развития науки проблемы, бихевиористы предложили рассматривать головной мозг человека, генерирующий сознание, как своего рода «черный ящик». Исследователи договариваются не задавать вопросов, что происходит там, внутри этой области. Точнее, они предполагают, что там на основе материальных нейрофизиологических процессов каким-то образом возникает сознание, но как именно это происходит – данный вопрос оказывается вне рассмотрения, подвергнутый своеобразной бихевиористской редукции. Разумность человека ученые-бихевиористы предлагают оценивать исключительно исходя из его поведения. На «вход» «черного ящика» подаются определенные сигналы, на выходе имеется определенная реакция. Совокупность типичных реакций на различные ситуации-стимулы, подаваемые на «вход» системы при наблюдении за поведением человека, признается характеристикой разумности. Такая элегантная в своей простоте стратегия позволяет обойти стороной все тяжеловесные метафизические споры о том, что, собственно, творится внутри «ящика», как именно возникает и функционирует сознание, что оно из себя представляет.

Очевидно, что Тьюринг применил тот же методический ход но уже по отношению к вопросу о разумности системы искусственного интеллекта. Не будем спрашивать, как именно технический материальный носитель может генерировать что-то подобное сознанию, разуму, ин-

теллекту, не будем спрашивать и о самих сущностных характеристиках разума. Представим машину в качестве «черного ящика» и будем оценивать ее степень разумности, основываясь лишь на исследовании ее поведения. Если в ответах на предлагаемые ей вопросы она окажется неотличимой от человека, будем считать машину (искусственно созданную техническую систему) разумной, интеллектуальной.

Однако, несмотря на столь подкупающую простоту для ученого, который не желает себя обременять излишними философскими спекуляциями, тест Тьюринга, для того, чтобы быть работоспособным, обязательно должен быть дополнен еще одним важным критерием. Необходимо задать те параметры, по которым он будет проводиться. Какие именно человеческие способности должна сымитировать искусственная техническая система в игре-имитации, чтобы ее ответы были признаны разумными? Для надлежащей характеристики данного критерия нам бы хотелось здесь ввести понятие «интервал Тьюринга», которое мы заимствуем у В.И. Моисеева [2]. В.И. Моисеев использовал очень удачный, емкий термин, точно выражающий специфику обсуждаемого критерия. Для того, чтобы тест Тьюринга заработал необходимо ввести тот интервал, в рамках которого он будет экзаменовать машину на предмет подобия ее поведения поведению человека.

Интервал Тьюринга – это совокупность вопросов, предлагаемых в тесте, вопросов, которые сформулированы, исходя из определенной парадигмы представления о разумности. Например, если мы в качестве первичного, интуитивно данного представления о разумности придерживаемся понятия «математическое вычисление» как важнейшего свойства, которое должно проявить себя здесь, то и вопросы будут выбраны нами соответствующие. В самом простом варианте это будет тривиальный математический тест. Мы спросим респондентов: «Сколько будет 2x2?», «Сколько будет 3428x6772?». Ясно, что имеющиеся на сегодняшний день машины будут способны пройти этот тест. Более того, если техническое устройство и «выдаст себя» по сравнению с человеком, то это обстоятельство, скорее всего, пойдет в зачет как раз машине, а не человеку. Ибо, очевидно, что современное техническое вычислительное устройство сможет выдать ответ на вопрос «Сколько будет 3428x6772?» быстрее своего биологического собрата. Парадигма, использующая понятие математического вычисления может быть расширена до исчисления логических понятий. Известно, что средствами современной логики мы можем формализовать процессы весьма сложных рассуждений, представленных в естественном языке. Кроме того, определенные логические методы (таблицы истинности, интерпретация тавтологичности формулы, построенной на языке логики высказываний)

позволяют нам оценивать истинность и ложность сложных рассуждений в зависимости от истинности и ложности входящих в них составных элементов. Все это приводит к тому, что соответствующим образом запрограммированная машина сможет воспроизвести сложное логическое рассуждение и оценить его истинность или ложность. Например, мы можем использовать в тесте Тьюринга типичные задачи из учебников по логике высказываний:

Кто-то (Иван, Петр, Алексей, Николай или Борис) съел банку варенья. Известно, что если съел Борис, то вместе с ним ели Иван и Николай. Если же съел Иван, то вместе с Петром. Петр и Алексей не могли есть варенье вместе, это мог быть только один из них. Алексей мог есть варенье только вместе с Николаем. По крайней мере, Николай или Борис съели варенье. Кто съел варенье?

Путем формализации и построения истинностных таблиц мы способны дать однозначный логический ответ на поставленный вопрос. Но дело в том, что точно такой же ответ будет способна дать и машина, ибо метод решения данной задачи, в соответствии с постулатами современной символической логики, не содержит ничего, превышающего функции математического вычисления, возможность приписывания которого «думающим машинам» нами на сегодняшний момент под сомнение уже не ставится. Следовательно, если использовать парадигму понимания интеллекта как исчисления понятий, то современный компьютер не испытает трудностей с прохождением теста Тьюринга.

Однако, будут те, кто не согласится, но не с результатами такого теста, а с самим «интервалом», который в нем выставлен. Если в качестве определяющего признака разумности принимать возможность распознавания объектов внешнего мира, оперирования с ними, подведения их под родовые понятия, то пройти тест Тьюринга машине окажется значительно сложнее. Здесь на вычислительную техническую систему придется «навешивать» аппаратуру из области робототехники – светоэлементы, отвечающие за внешнее восприятие, и обширный буфер постоянной памяти, хранящей образцы возможных объектов. Но в принципе, даже первое техническое устройство, имитирующее восприятие, сконструированное когда-то Ф. Розенблантом [3. 18], способно пройти простейший тест Тьюринга по интервалу, заданному парадигмой «интеллект как восприятие». Респондентам может быть показана буква «А» и задан вопрос типа: «Какую букву вы сейчас видите?» И человек, и перцептрон Розенбланта способны дать одинаковый ответ: «Первую букву английского алфавита». Другое дело, что возможности подобных устройств из области робототехники, по-прежнему, остаются крайне ограниченными по сравнению с вычислительными машинами. Если

перцептрону показать торчащий из-за кресла кошачий хвост и задать вопрос «Что это такое?», то пройти тест Тьюринга данному техническому устройству уже не удастся.

Есть еще более радикальный аргумент со стороны тех, кто будет препятствовать представлению о разумных машинах, и о нем упоминает сам Тьюринг: «Этот аргумент прекрасно выражен в Листерской Речи профессора Джефферсона в 1949 году: ‘Пока машина не напишет сонета или концерта, вдохновленного чувствами, а не полученного в результате случайного сочетания символов, мы не сможем согласиться с тем, что машина равна мозгу. Никакой механизм не может почувствовать (а не просто показать при помощи какого-либо несложного ухищрения) удовлетворения от удачи, печали от того, что его контакты перегорели, испытать удовольствие от похвалы, расстроиться из-за своих ошибок, быть очарован сексом, сердиться или впадать в депрессию, когда он не может получить желаемого’» [4. 52]. Возможно, такую позицию следовало бы трактовать как парадигму «интеллект как внутреннее чувство», однако подобный редукционизм кажется уже слишком абсурдным, ибо разум здесь просто превращается в эмоциональную жизнь, мы просто устраняем одно понятие, заменяя его другим.

Пожалуй, что данная позиция, делающая акцент на внутреннем мире, на чувственности человека, с наименьшими потерями для понятия разумности, интеллектуальности, могла бы быть встроена в такие парадигмы, как «интеллект как интенциональность» и «интеллект как самоидентичность». Первая из этих парадигм, сформулированная, прежде всего, в известной статье Д. Серла [5] настаивает на том, что основным критерием разумности должна выступить интенциональность – способность человека отдавать себе отчет о содержании своих психических состояний, понимать с чем именно он сейчас имеет дело. Серл считает, что машина никогда не будет способна на продуцирование подобной характеристики. Парадигма «интеллект как самоидентичность», о которой упоминает, в частности, Л. Хосер [6], утверждает, что отличительным признаком разумности следует считать способность человека отдать отчет о себе самом, сказать себе «я».

Нельзя сказать, что машина, выполняющая тест Тьюринга с интервалом, заданным последними двумя парадигмами, потерпела бы неудачу. Машина могла быть снабжена подходящей программой, которая позволила бы ей умело управляться со словом «я» и со всеми перформативными высказываниями, призванными позиционировать субъективную жизнь сознания. И тест не показал бы разницы между ответами человека и искусственной технической системой.

Однако подобные результаты теста нас, конечно же, не убеждают. Скорее, мы должны признать, что с точки зрения последних двух парадигм в качестве основной характеристики разумности признается нечто такое, что вообще не может быть использовано в качестве «интервала Тьюринга». Ведь интервал, исходя из первоначального замысла теста, должен выставляться по такому свойству, которое в принципе может быть распознано в поведенческих речевых реакциях. Но акцент на внутреннюю жизнь, на интенциональность, на самоидентичность – это акцент на те свойства, которые не удовлетворяют данному критерию вообще. Внутренняя жизнь в принципе «не прозрачна». Внешнее поведение о ней ничего не может сказать. Никогда нельзя утверждать наверняка, стоит ли за выраженными словами подлинное интенциональное переживание субъекта или нет. Поэтому с точки зрения данных парадигм понимания разума тест Тьюринга вообще оказывается непригодным для выполнения поставленной задачи – фиксации интеллектуальных действий.

Однако критические аргументы в данном случае могут быть выдвинуты не только по отношению к тесту, но и по отношению к самой философской позиции, со стороны которой высказываются сомнения в реализации замысла Тьюринга. Вышеуказанная позиция тоже является достаточно уязвимой, и важнейший контраргумент здесь, кстати, упоминаемый уже и самим автором теста, состоит в следующем. Отрицание возможности применения теста Тьюринга касаемо характеристик интенциональности и самоидентичности ставит нас в тупик не только перед вопросом «Мыслит ли машина?», но и перед вопросом «Мыслит ли человек?». Данные парадигмы понимания разумности базируются на радикальном солипсизме, который, как известно, никогда не был в особом почете в среде эпистемологов в виду банальной непродуктивности этой позиции. Если сказать, что обнаружить существенные свойства разумности субъект может только в отношении своей собственной психической жизни, то невозможно будет доказать интеллектуальность действий не только искусственной технической системы, но и системы естественно-биологической, т.е. человека. По сути, наше предположение, что за словами, которые слетают с уст такой биологической системы как человек, стоит внутренняя психическая жизнь, есть только лишь предположение, и оно, в данном случае, является гипостазированием психического не в меньшей степени, чем при приписывании субъективности технической системе.

Обсуждение различных парадигм понимания разумности в отношении теста Тьюринга может быть продолжено (мы рассмотрели здесь еще не все значимые воззрения; например, требуют особого исследова-

ния такие парадигмы, как «интеллект как рефлексия» и «интеллект как креативность»), однако цель настоящей статьи состояла не в этом. Скорее, наш замысел заключался в том, чтобы обратить внимание на определенную теоретическую наивность теста Тьюринга, которая обнаруживается в представленном обсуждении парадигм.

Чего именно хотел добиться автор данной методической разработки? Он намеревался получить надежный способ фиксации разумного поведения той или иной системы. Если взять только этот аспект назначения теста, то можно констатировать, что его проект, по большей части, удался (хотя тоже не во всех случаях, если вспомнить парадигмы, где тест не действует вообще). Но ведь Тьюринг, как кажется, хотел большего. Он хотел прекратить споры о сущности разума, подразумевая, что его тест будет способен фиксировать разумное поведение таким образом, чтобы не задавать вопросы об основных характеристиках самой разумности, не отвечать на вопрос «Что считать разумным?» И относительно такого разворота проблемы нам следует уяснить, что его проект потерпел неудачу. Ответ на вопрос «Что считать разумным?» уже заранее предполагается при проведении теста, ибо он должен контролировать тот «интервал Тьюринга», который будет применен в том или ином варианте эксперимента. Такой «интервал» является критерием, соответствие/несоответствие которому позволит фиксировать разумность поведения системы. И игнорировать этот критерий никак нельзя, он является условием возможности эксперимента, без которого тест просто не запустится. А это значит, что предложенный одним из родоначальников исследований в области искусственного интеллекта методический ход не приносит желанного освобождения от философских споров о сущности разума, напротив, оказывается, что выбор той или иной философской парадигмы понимания разумности является отправной точкой бихевиористского теста.

ЛИТЕРАТУРА

1. *Turing, A.M.* Computing machinery and intelligence. Mind, № LIX, 1950. P. 433-460.
2. *Моисеев, В.И.* Интервал Тьюринга и имитация интеллекта // Философия искусственного интеллекта. Материалы Всероссийской междисциплинарной конференции. М.: ИФ РАН, 2005. С. 307-309.
3. *Ладов, В.А.* Философские проблемы искусственного интеллекта: Учебно-методическое пособие. Томск, 2005.
4. *Тьюринг, А.М.* Вычислительные машины и разум. // Хофштадтер, Д.; Деннет, Д. Глаз разума. Самара, 2003.
5. *Серл, Д.* Мозг, сознание и программы. // Аналитическая философия: Становление и развитие (антология). М., 1998. С. 376-400.
6. *Hauser, L.* Why Isn't My Pocket Calculator a Thinking Thing? // Minds and Machines, Vol. 3, No. 1 (February), 1993.