АННА ГЕННАДЬЕВНА ГОРБАЧЁВА

Аспирант кафедры философии
Новосибирского государственного университета
экономики и управления.

E-mail: gorbacheva.a.g@gmail.com

УДК 004.896

ТЕСТ ТЬЮРИНГА: ДИАГНОСТИКА ЧЕЛОВЕЧЕСКОГО В ИНТЕРФЕЙСЕ «ЧЕЛОВЕК — МАШИНА»

работа выполнена при поддержке Российского научного фонда в рамках проекта «Построение неклассической антропологии. Новая онтология человека» (грантовое соглашение № 14-18-03087)

Тест Тьюринга рассмотрен с учетом стремительного развития компьютерных и сетевых технологий. Введены понятия контекстного искусственного интеллекта, искусственного интеллекта определенного уровня и абсолютного искусственного интеллекта. Приведены аргументы в пользу того, что машина сможет пройти тест не только из-за своего совершенствования, но и благодаря изменению взаимодействия между человеком и машиной (аутсорсинга интеллектуальных функций человека) вместе со снижением уровня человеческого интеллекта. Спрогнозированы возможные последствия создания абсолютного искусственного интеллекта и предложен тест для него.

Ключевые слова: искусственный интеллект, тест Тьюринга, «китайская комната», жизненный аутсорсинг.

Anna G. Gorbacheva

TURING TEST: DIAGNOSTICS OF THE HUMAN ESSENCE IN THE MAN-MACHINE INTERFACE

The paper is devoted to analysis of the problem which is connected with intensification of «man-machine» interaction and contemporary trends of life outsourcing, when many human intrinsic functions are delegated to computers. This problem is considered by looking at the Turing

test over a prism of contemporary technologies. We introduce notions of context artificial intelligence, a level of artificial intelligence and absolute artificial intelligence. We argue in favor of the fact that potentially a machine will be able to pass the Turing test not only due to its modernization, but also due to changing of the «machine-man» interaction and weakening of human intelligence. We try to predict some consequences of (hypothetical) creating the absolute artificial intelligence and introduce a test for it.

Keywords: artificial intelligence, Turing test, Chinese room, life outsourcing.

Проблема создания искусственного интеллекта вызывает устойчивый интерес как у специалистов в области информатики, так и у философов, пытающихся ответить на вопрос «Может ли машина мыслить?». Основоположником данного направления исследований является А. Тьюринг, предположивший, что через 50 лет после его публикации (в 2000 г.) мыслящая машина будет создана [Turing A. 1950] . Однако, несмотря на усилия ученых, до сих пор такая машина не только не создана, но даже нет убедительных аргументов в пользу того, что ее в принципе можно создать. Более того, Д. Серль предложил опровержение данной гипотезы в виде концепции так называемой «китайской комнаты» [Searle J. 1980]. Он утверждал, что машина способна пройти тест Тьюринга, даже не «умея» мыслить, т. е. не понимая семантики языка. Ей достаточно уметь оперировать с синтаксисом: поверхностными грамматическими структурами и правилами их построения.

Тест, предложенный Тьюрингом, — это мысленный эксперимент, заключающийся в том, что некий собеседник-исследователь, взаимодействуя либо с машиной, претендующей на обладание искусственным интеллектом, либо с человеком, должен определить, с кем из них он взаимодействует. Согласно гипотезе Тьюринга, машина, которую собеседник-исследователь не сможет отличить от человека, обладает искусственным интеллектом.

Серль предложил следующий эксперимент. В некоторую комнату помещают человека, не знающего китайского языка, но обладающего инструкцией с описанием правил составления ответов на вопросы, записанные на китайском языке. Серль утверждал, что такой человек, руководствуясь инструкцией, сможет составлять адекватные ответы, не понимая смысла ни вопросов, ни ответов.

Несмотря на критику теста Тьюринга со стороны Серля и его последователей, интерес к проблеме создания машины, успешно проходящей данный тест, попрежнему высок. Создание такой машины оказывается интересным уже само по себе, несмотря на разногласия относительно того, можно ли считать успешное прохождение теста свидетельством наличия интеллекта у машины [Abramson D. 2011].

Целью работы является анализ концепций Тьюринга и Серля с современной позиции, учитывая стремительное развитие сетевых и компьютерных технологий, способствовавших не просто автоматизации рутинных процессов, но и передаче целого ряда человеческих функций машинам. В работе С. А. Смирнова [Смирнов С. А. 2012] это явление названо «жизненным аутсорсингом».

«Китайская комната» и человек

Основная позиция, с которой мы рассмотрим тест Тьюринга, — это его практическая значимость. Нас будут интересовать именно внешние проявления тестируемой машины, а не наличие или отсутствие у нее сознания и понимания вопросов и ответов. Выражаясь языком Серля, для нас не так важно, находится ли машина в «китайской комнате» или нет. Более того, на наш взгляд, сами люди за-

частую оказываются в ситуациях сродни «Китайской комнате», не отличаясь в этом смысле от машины, которая не понимает семантику языка. Рассмотрим три примера, иллюстрирующие данную идею.

Вначале приведем описание концепции «китайской комнаты» Серля [Searle J. 1980]. Человека, не знающего китайского языка, помещают в комнату с окошком. Этот человек не просто не знает китайского языка, но и не может даже отличить китайского от японского или другого языка, основанного на иероглифах. Человек должен письменно на китайском языке отвечать на вопросы, также заданные на китайском. Для этой цели человека предварительно снабжают правилами манипуляции китайскими иероглифами, причем эти правила написаны на его родном языке. Правила являются не словарем, а неким набором алгоритмов для конструирования ответов. Серль утверждает, что человек, пользуясь выданными ему правилами, может составлять вполне адекватные ответы на вопросы и тем самым пройти тест Тьюринга, совершенно не понимая семантики (смысла) входного текста и своих ответов; он будет только манипулировать синтаксисом. Серль утверждал, что если и удастся создать машину, проходящую тест Тьюринга, то она будет его проходить только на уровне синтаксиса, не понимая и не осознавая семантики.

В первом примере речь пойдет о студенте, готовящем реферат. В настоящее время, когда Интернет предоставляет доступ почти к любой информации, подготовка реферата может свестись к поиску нужной информации в Сети с последующим копированием и вставкой найденных текста и иллюстраций; по большому счету, необязательным может оказаться даже понимание того, о чем идет речь в копируемом тексте. Наверняка многие преподаватели сталкивались с ситуацией, когда студент приносит в целом неплохую работу, но на вопросы по ней ответить не может; однако если ему задать вопрос и дать некоторое время на подготовку, то он, скорее всего, найдет ответ на него. Таким образом, студент может подготовить качественный реферат практически на любую тему (и даже ответить на вопросы по нему), не слишком в ней разбираясь. Тем самым студент как бы «помещает» себя в «китайскую комнату», не понимая семантики своего реферата.

Второй пример тесно связан с первым, но имеет более длинную историю: речь идет о списывании. Предположим, что школьник сдает списанное домашнее задание, и учитель, не проводя каких-либо экспериментов (не сравнивает работу с другими, не задает вопросы по теме, не пользуется информацией о прошлых заслугах ученика), будет считать, что задание выполнено.

Примеры показывают, что уже сейчас люди могут успешно выполнять поставленные задачи, не понимая их смысла (семантики).

В настоящее время ситуация «китайской комнаты» становится все более распространенной. Под влиянием коммерциализации и агрессивного маркетинга сама жизнь заставляет людей действовать шаблонно, не задумываясь о смысле и значении своих действий и слов.

По этой причине сами концепции Серля и Тьюринга не выводят нас на новое решение проблемы — это проблема выходит за рамки теста и за рамки собственно технического устройства. Проблема теперь заключается не в том, чтобы сделать машину умной (здесь достижения компьютерных технологий и биоинженерии все более впечатляющи), а в том, что во взаимодействии человека и машины (где изначально машина играла роль исключительно некоего умного инструмента) происходит перераспределение базовых функций.

В тесте Тьюринга испытуемым объектом является машина, «старающаяся» пройти тест, и гипотеза Тьюринга заключалась в том, что через 50 лет после его

публикации машина сможет это сделать. Это означает, что в понимании Тьюринга искусственный интеллект — это система, «интеллект» которой разработчики пытаются «поднять» до уровня человеческого. Чтобы машина смогла пройти тест Тьюринга, разработчики создают все более и более изощренные программы, подразумевая тем самым, что именно интеллект машины требует совершенствования. Однако очевидно, что если в качестве собеседника-исследователя выбрать человека с низким интеллектом, неспособного задавать каверзные вопросы, или просто человека в состоянии наркотического опьянения, то он наверняка не сможет отличить от человека даже машину с относительно простой программой. Значит, машина может успешно пройти тест не только тогда, когда ее интеллект поднимется до человеческого уровня, но и когда собеседник-исследователь будет обладать низким интеллектом (в работе [Горбачёва 2014], например, приведены аргументы, что в будущем возможно снижение интеллекта среднего человека).

Таким образом, довольно сложно считать постановку теста Тьюринга корректной, особенно в условиях, когда современный человек все больше своих интеллектуальных функций передает машинам. При этом вообще теряется базовое различие — где здесь машина и где человек? Далее мы введем понятия контекстного искусственного интеллекта, интеллекта с определенным уровнем и абсолютного искусственного интеллекта.

Практическая и теоретическая значимость теста Тьюринга и проблема искусственного интеллекта

Тьюринг задал следующий вопрос: «Может ли машина мыслить?». Серль уточнил его: «Могут ли машины иметь осознанные мысли в таком же смысле, что и мы с вами?». Но корректна ли такая постановка вопроса? На наш взгляд, не совсем. Здесь мы придерживаемся мнения Т. Найджела, который утверждал, что человек не может в полной мере представить себе опыт другого существа (летучей мыши), поскольку другие существа имеют совершенно другие способы восприятия окружающего мира [Nagel T. 1974]. Но усомниться в том, что другие живые существа абсолютно обделены интеллектом, безусловно, нельзя. Значит, даже если машина и будет мыслить, скорее всего, она не будет это делать так же, как люди.

На наш взгляд, значимость «умения» машины успешно пройти тест Тьюринга существенно выше ее «умения» понимать семантику вопросов. Другими словами, для практики не так важно, находится ли собеседник в «китайской комнате» или нет — главное, чтобы он вел себя так, как требуется. В конце концов, искусственный интеллект создается не только из чисто теоретического интереса, но и с целью принести пользу обществу, что, безусловно, является чисто практическим интересом. Такого мнения придерживается В.П. Литвинов, утверждающий, что вопрос существования искусственного интеллекта Тьюринг ставил «не просто как познавательный, а как конструктивно-технический» [Литвинов В. П. 2012].

Здесь важно ответить на следующий вопрос. Что же мы хотим от машины — проверить ее интеллект или отличить от человека? Все-таки человека человеком делает не только интеллект, но и, например, эмоции. Дж. Мегиль приводит ряд аргументов в пользу того, что, с одной стороны, эмоции являются важной частью познавательной деятельности человека, а с другой — машина не обязана их испытывать, чтобы осуществлять соответствующие виды когнитивной и, следовательно, интеллектуальной деятельности [Megill J. 2014]. Утверждается, что если машина и проиграет человеку в интеллектуальном соревновании, то это произойдет не по причине отсутствия эмоций. Таким образом, мы можем сделать следующие выводы:

1. Для успешного прохождения теста Тьюринга машина должна уметь реагировать на эмоции и, возможно, имитировать их.

2. Тот факт, что машина не переживает эти эмоции в действительности, не препятствует тому, что машина может осуществлять интеллектуальную деятельность.

3. Человек, скорее всего, не сможет понять, что чувствует другой организм (в данном случае — машина) в таком же виде, как он это в действительности делает.

Значит, важным фактором успешного прохождения теста Тьюринга машиной является ее умение понимать и имитировать (но не обязательно переживать) человеческие эмоции. Например, машина должна уметь реагировать на эмоциональные фразы, сказанные человеком.

Рассмотрим, как передаются эмоции в известном формате CME (computer mediated environment). Для этого используются смайлики, эмоционально украшенные слова, знаки препинания. Причем набор всех этих знаковых средств довольно ограничен и легко может быть занесен в машину вместе с возможными реакциями на них. При этом отсутствует интонация, взгляды, прикосновения и прочие невербальные формы коммуникации, позволяющие проявлять эмоции при живом общении.

Люди перенесли значительную часть своих коммуникаций в Сеть, как правило, в письменную или визуальную форму (короткие сообщения, электронная почта, чаты, фото- и видеоматериалы). Даже доступный Skype используется далеко не всегда. Люди проявляют и получают эмоции, передающиеся посредством весьма ограниченного набора. Люди все меньше времени проводят в живом общении, которое, без сомнения, намного более эмоционально насыщено, чем коммуникации через CME. Вероятно, проводя много времени за виртуальным общением, люди эмоционально беднеют, становясь неспособными различать тонкие эмоции, которые невозможно передать текстом. Следовательно, они и в живом общении будут более ограничены.

Отсюда следует, что машине будет значительно легче имитировать чувства эмоционально бедного человека и убедить его в том, что машина – это человек. С развитием компьютерных технологий процент эмоционально бедных людей будет расти и, значит, процент людей, эмоции которых можно имитировать, также будет расти.

Л. В. Мурейко говорит о деперсонификации людей, действия которых отличаются высокой степенью механистичности и автоматизма, присущего машинам [Мурейко Л. В. 2009]. Само человеческое существование становится во многом стандартизированным, а значит программируемым.

Н. Л. Караваев отмечает, что информационные технологии ограничивают сознание человека, фиксируя его мышление определенными рамками [Караваев Н. Л., Окулов С. М. 2011]. Человек все меньше и меньше ищет новые решения проблем, используя уже существующие методы. Караваев пишет, что человек теряет интуитивное понимание мира, стараясь увидеть во всем структуру. Но ведь именно наличие способностей увидеть какие-то нестандартные решения и отличает человека от машины. Поэтому здесь люди приближаются к машинам: «Мы превращаемся в такие же машины» [Там же].

И. Ю. Алексеева говорит о схожей тенденции, заключающейся в том, что необходимость быстрого извлечения информации оставляет в стороне вопрос развития интеллектуальных способностей человека [Алексеева И. Ю. 2012]. Н. Л. Караваев утверждает, что первые информационные революции (язык и письменность)

дали толчок к развитию интеллекта человека, а последние несут в себе больше негативных последствий [Караваев Н. Л. 2013].

Контекстный искусственный интеллект

В данной работе мы предложим понятие контекстного искусственного интеллекта и обоснуем, что он уже создан и имеет важное практическое значение.

Будем говорить, что машина обладает контекстным искусственным интеллектом, если она успешно проходит тест Тьюринга в конкретном контексте и непредвзятом исследовании, но может не пройти тест, если хотя бы одно из этих условий не выполнено. Под контекстом будем понимать комбинацию цели взаимодействия и правил (интерфейса) взаимодействия.

В своем исходном виде тест Тьюринга не предполагал никаких ограничений на вопросы, которые может задавать собеседник-исследователь. Следовательно, вопросы могли быть самыми каверзными и неожиданными. Однако к настоящему времени не создано машин, которые бы могли успешно пройти такой идеальный тест; более того, до сих пор не предложено убедительных доказательств принципиальной возможности создания такой машины. Одной из причин данной проблемы, на наш взгляд, является неточная постановка задачи. Об этом уже говорилось выше: во-первых, в исходном тесте Тьюринга нет конкретных предположений относительно интеллектуальных возможностей собеседника-исследователя, и, во-вторых, нет гарантий того, что интеллект человека (какого бы то ни было: среднего, представителя элит, с образованием, без образования) является постоянной величиной. Таким образом, чтобы хоть как-то ответить на вопрос о возможности создания искусственного интеллекта, можно наложить некоторые разумные ограничения на условия, в которых проводится тестирование, тем более, что, как подчеркивалось выше, искусственный интеллект в данной статье рассматривается прежде всего с практической точки зрения.

Рассмотрим пример: покупку товаров через автоматизированный терминал. Это типичный пример контекстного искусственного интеллекта. В данном случае совершенно точно ясна цель взаимодействия – покупка товара и определен интерфейс, через который происходит взаимодействие пользователя с компьютером. Покупатель взаимодействует с компьютером посредством выбора товара нажатием кнопок и передачей денежных средств. Если покупатель будет взаимодействовать с продавцом через такой же интерфейс, то отличить компьютер от человека будет невозможно.

Теперь рассмотрим пример выхода за рамки контекста. Предположим, что истинная цель покупателя состоит не в покупке товара, а в выяснении того, компьютер это или человек. Здесь исследователь предвзят, и его цель не соответствует цели создания данного терминала. В таком случае он может взломать терминал или спросить у информированных лиц о том, что в нем в действительности скрыто. В данном случае покупатель нарушил цель создания контекста. Если человек решит расплатиться через терминал не положенным образом, а, например, с помощью карты, то он нарушит интерфейс. Если помимо покупки товара пользователь захочет поговорить с продавцом на отвлеченные темы, как часто делают в магазинах постоянные клиенты, то он также выйдет за рамки контекста, изменив цель.

Интеллект определенного уровня

Второе понятие, которое мы введем, — это искусственный интеллект определенного уровня.

Несмотря на то, что к настоящему моменту нет строгих неопровержимых доказательств наличия различного уровня интеллекта у разных людей, существует некая интуитивная уверенность в этом факте. Разработаны способы измерения интеллектуального уровня, например традиционные оценки или тест IQ. Абсолютного доверия к ним нет, но они все же имеют довольно сильную связь с интуитивно понимаемым уровнем интеллекта. Таким образом, можно выдвинуть две следующие гипотезы:

1. Уровень интеллекта у разных людей разный.

2. Способ измерения уровня интеллекта должен существовать, хотя он еще и не открыт.

В рамках этих гипотез можно ввести способ измерения уровня искусственного интеллекта. Искусственный интеллект уровня *X* — это такая система, которая проходит тест Тьюринга, если исследователь имеет интеллект уровня *X*. Искусственный интеллект определенного уровня можно использоваться как тест человеческого интеллекта — человек имеет такой уровень интеллекта, какой уровень искусственного интеллекта он в состоянии распознать.

Абсолютный искусственный интеллект

Под абсолютным искусственным интеллектом мы будем понимать интеллект, который невозможно отличить от человеческого ни при каких условиях. Хотя здесь и возникает некоторая двусмысленность, заключающаяся в том, что же следует понимать под неотличимостью от человеческого интеллекта и под самим человеческим интеллектом, а точнее под уровнем человеческого интеллекта. Можно сказать, что абсолютный искусственный интеллект — это такой интеллект, который проходит тест Тьюринга при любом собеседнике-исследователе, однако нет гарантии, что он будет являться сильным искусственным интеллектом по Серлю, т. е. понимать семантику вопросов. Поэтому, на наш взгляд, необходимо отталкиваться именно от того, что этот интеллект не должен отличаться от человеческого.

Абсолютный искусственный интеллект должен уметь синтезировать знания и направлять их в пользу выживания и приобретения власти. Следовательно, очевидным тестом на абсолютный искусственный интеллект может быть тот факт, что машина будет непредсказуемым образом затруднять человеку задачу «нажимать кнопку», чтобы не дать ему отключить себя. Под непредсказуемостью мы понимаем то, что человек не будет влиять на принятие решения об отключении. Сейчас во многих системах человек действительно не может отключить компьютер, например, на атомных станциях, *осознавая*, что это приведет к катастрофическим последствиям, или же сами разработчики создают степени защиты от выключения. Мы говорим не об этом. Здесь человек не отключает приборы по своей инициативе: либо просто осознавая, что этого делать нельзя из-за нежелательных последствий, либо потому, что другие люди спроектировали систему так, чтобы ее было трудно отключить. Мы же говорим о том, что даже если человек захочет отключить машину, он не сможет этого сделать из-за ее непредсказуемых действий. Причем здесь, строго говоря, не понадобится абсолютный интеллект. Достаточно будет только интеллекта, чей уровень будет выше, чем у любого живущего на земле.

Мы приходим к выводу, что абсолютный искусственный интеллект неотделим от физической реальности. Другими словами, машина должна иметь конкретные средства взаимодействия с окружающей средой; и наиболее очевидным воплощением искусственного интеллекта могут стать роботы, точнее, киборги,

интегрирующие в себе человека и машину и фактически нивелирующие взаимодействия между человеком и машиной, вмещая это взаимодействие внутрь себя.

В настоящей статье приведены доводы в пользу актуальности всестороннего анализа теста Тьюринга на сегодняшний день. Проведен краткий критический анализ концепции «китайской комнаты» Серля, учитывая стремительное развитие современных компьютерных и сетевых технологий. Введены понятия контекстного искусственного интеллекта, искусственного интеллекта определенного уровня и абсолютного искусственного интеллекта. Приведены аргументы в пользу того, что машина сможет пройти тест Тьюринга не только благодаря своему совершенствованию, но и благодаря изменению процесса взаимодействия человека и машины (аутсорсинга интеллектуальных функций человека) вкупе с возможным снижением человеческого интеллекта в будущем. Спрогнозированы возможные последствия создания абсолютного искусственного интеллекта и предложен тест для него.

Литература

Алексеева И. Ю. 2012. Информационная компетентность, естественный интеллект и НБИКС-революция // Информ. общество. — 2012. — № 5. — С. 9–15.

Горбачёва А. Г. 2014. Человеческий интеллект: возможные изменения под влиянием информационных технологий и высокотехнологичных устройств // Идеи и идеалы. — 2014. — № 1. — Т. 2. — С. 135–142.

Караваев Н. Л. 2013. Об антропологических проблемах информационного общества // Философские проблемы информационных технологий и киберпространства. — 2013. — № 1. — С. 65–73.

Караваев Н. Л., Окулов С. М. 2011.Информационно-коммуникационные технологии и человек // Вестн. Вят. гос. гуманитарн. ун-та. — 2011. — № 4. — С. 47–51.

Литвинов В. П. 2012. Актуальность задачи Тьюринга (Can machines think?) // Философ. проблемы информ. технологий и киберпространства. — 2012. — № 1. — С. 93–100.

Мурейко Л. В. 2009. О природе массового сознания в контексте исследований «искусственного интеллекта» // Изв. рос. гос. пед. ун-та им. А. И. Герцена. — 2009. — № 110. — С. 90–100.

Смирнов С. А. 2012. Фармацевтика антропологических трендов. Антропологический форсайт // Вестн. НГУЭУ. — 2012. — № 1. — С. 88–104.

Turing A. 1950. Computing machinery and intelligence // Mind. — 1950. — No. 59. — Pp. 433–460.

Searle J. 1980. Minds, brains, and programs // The behavioral and brain sciences. — 1980. — № 3. — Pp. 417–457.

Abramson D. 2011. Philosophy of mind is (in part) philosophy of computer science // Minds & Machines. — 2011. — № 21. — Pp. 203–219.

Nagel T. 1974. What is it like to be a bat? // The Philosophical Review. — 1974. — Vol. 83. — № 4. — Pp. 435–450.

Megill J. 2014. Emotion, cognition and artificial intelligence // Minds & Machines. — 2014. — № 24. — Pp. 189–199.