ISSN 2524-2369 (Print)
ISSN 2524-2377 (Online)
УДК 004.8:7.01
https://doi. org/10.29235/2524-2369-2022-67-1-15-20

Поступила в редакцию 14.05.2021
Received 14.05.2021

И. К. Ставровский

Институт философии Национальной академии наук Беларуси, Минск, Беларусь

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ТВОРЧЕСТВО

Аннотация. Одной из самых знаменитых альтернатив теста Тьюринга является тест Лавлейс, который предлагает использовать понятие «творчество» как способ оценки способности искусственного интеллекта мыслить в том же смысле, что и человек. В статье демонстрируется, что понятие «творчество» слишком неоднозначно, чтобы использовать его как критерий чего-либо. Кроме того, показано, что обе версии теста Лавлейс, по сути, являются бихевиористскими, поэтому принципиально не могут доказать способность машины мыслить.

Ключевые слова: искусственный интеллект, тест Тьюринга, тест Лавлейс, психология творчества, философия творчества

Для цитирования: Ставровский, И. К. Искусственный интеллект и творчество / И. К. Ставровский // Вес. Нац. акад. навук Беларусі. Сер. гуманітар. навук. – 2022. – Т. 67, № 1. – С. 15–20. https://doi. org/ 10.29235/2524-2369-2022-67-1-15-20

Igor K. Stavrovsky

Institute of Philosophy of the National Academy of Sciences of Belarus, Minsk, Belarus

ARTIFICIAL INTELLIGENCE AND CREATIVITY

Abstract. The Lovelace Test is one of the most famous alternatives to the Turing Test. It suggests to use the concept of creativity as a way to estimate the ability of artificial intelligence to think in the same sense as a human being. It is demonstrated in the article that the concept of creativity is too ambiguous to be used as a criteria of anything. It is shown that both versions of the Lovelace Test are inherently behaviorist and therefore cannot prove the ability of a machine to think.

Keywords: artificial intelligence, Turing test, Lovelace test, psychology of creativity, philosophy of creativity

For citation: Stavrovsky I. K. Artificial intelligence and creativity. *Vestsi Natsyyanal'nai akademii navuk Belarusi. Seryia humanitarnykh navuk = Proceedings of the National Academy of Sciences of Belarus. Humanitarian Series,* 2022, vol. 67, no. 1, pp. 15–20 (in Russian). https://doi.org/10.29235/2524-2369-2022-67-1-15-20

Введение. Долгое время способности человеческого разума считались уникальными, единственными в своем роде. Машины, в свою очередь, рассматривались как устройства, способные выполнять только примитивные задачи, которые не сравнятся по сложности с тем, на что способен человек. Подобное представление зависело от уровня развития технологий прошлого, поэтому по мере развития вычислительной техники становилось все более очевидным, что исключительность человеческого разума была сильно переоценена.

Еще сравнительно недавно навык быстрого счета в уме, хорошая память или, к примеру, мастерство в шахматах рассматривались как надежные признаки высокого интеллекта, причем исключительно человеческого. Однако компьютеры доказали, что могут справляться со всем этим заметно лучше нас. Это должно было сильно пошатнуть или даже полностью разрушить представление людей об уникальности своих когнитивных способностей.

Отчасти это случилось, но гораздо интереснее другое. Каждый раз, когда компьютеры демонстрировали свое превосходство над человеком или даже просто овладевали «исконно человеческими» способностями, эти самые способности объявлялись чем-то тривиальным, механическим и не заслуживающим внимания. Проще говоря, престиж этих способностей падал в той степени, в которой ими овладевали компьютеры.

У данного события есть два взаимодополняющих следствия.

Первое заключается в том, что произошла демистификация этих способностей. Хотя они еще могут впечатлять, мы все же понимаем, что здесь нет никакой магии или тайны. Это можно

сравнить с тем, как компьютеры воспринимают люди, не разбирающиеся в их устройстве. Да, человек не знает, что именно происходит внутри машины, но он понимает, что это просто вопрос получения соответствующего знания.

Второе следствие объясняет описанную выше потерю престижа тем, что люди стремятся любой ценой сохранить свою уникальность. Ради этого можно «пожертвовать» некоторыми способностями, признав их не исключительно человеческими. Главное сохранить что-то все еще делающее нас особенными, а не всего лишь биороботами.

Таким образом, хотя люди пересмотрели представления о своей уникальности, они вовсе от нее не отказывались.

Одной из главных черт бытия человеком до сих пор считается способность к творчеству. Многие скажут: «Пусть машины занимаются скучными вычислениями, но им никогда не написать симфонию». Отчасти это утверждение справедливо. Так, в 2019 г. искусственный интеллект дописал неоконченную симфонию Густава Малера. Однако это не создание нового произведения с нуля. Машина анализировала произведения умершего композитора, выделяя общие для его творчества паттерны, на основании чего предлагала различные варианты продолжения симфонии. При этом выбор лучшего варианта осуществлял человек. Более того, поклонники композитора заметили переход от оригинального произведения к части, которая была создана компьютером. Они также обратили внимание, что в дописанной части симфонии пропала «эмоциональная глубина» [3].

Мы можем предположить, что мнение поклонников Малера было предвзятым. Однако это не отменяет того факта, что выбор продолжения симфонии осуществлял человек. Таким образом, искусственный интеллект выступает не в роли творца, а скорее в роли высокотехнологичного инструмента. Его существование – не более чем любопытный факт, который никак не оспаривает место человека в мире.

Тем не менее вопрос о том, может ли машина творить, не теряет своей актуальности. Хотя современные разработки не могут впечатлить нас своими творческими способностями, однако в будущем ситуация может сильно измениться. Но даже если машины никогда не смогут творить в том же смысле, что и человек, сам вопрос может помочь разобраться в понятии «творчество».

Две версии теста Лавлейс. Классическим методом оценки способности искусственного интеллекта демонстрировать разумное поведение (т. е. неотличимое от человеческого) является тест Тьюринга. Во время тестирования основной задачей искусственного интеллекта является победа в игре в имитацию, т. е. в ходе диалога убедить собеседника-человека в том, что он тоже человек. При этом собеседник-человек не видит и не слышит, с кем он разговаривает, поэтому он может полагаться только на текст, появляющийся на экране монитора.

Тест Тьюринга много раз подвергался критике со стороны разных авторов, однако мы не будем останавливаться на этом подробно. Достаточно указать, что для прохождения теста Тьюринга разработчики искусственного интеллекта используют мелкие уловки и трюки, даже не пытаясь создать разумную машину [1, с. 40].

В качестве альтернативы тесту Тьюринга Сельмер Брингсйорд, Пол Белло и Дэвид Ферруччи предложили тест Лавлейс [6], названный в честь леди Ады Лавлейс – математика и одного из разработчиков первой вычислительной машины. Лавлейс утверждала, что компьютер лишь тогда можно будет считать разумным, когда он сможет создавать новое.

Здесь важно отметить, что новизна не должна быть результатом ошибки или случайности. Действительно, нет ничего сложного в том, чтобы написать программу, которая будет создавать случайные последовательности чисел, слов, цветовых пятен, звуков и т. д. В некоторых случаях мы даже сможем придать смысл этим последовательностям, находить их красивыми и интересными. Тем не менее это будет лишь игрой случайности.

Это подводит нас к другому важному замечанию: новизна не должна сводиться к субъективному ощущению. Для маленького ребенка новым будет являться многое из того, что взрослым людям кажется тривиальным. Аналогично неспециалист в какой-либо области будет считать новым практически все, что он узнает, в то же время для эксперта практически все будет знакомо. Таким образом, нас интересует новизна не как психологический факт удивления того или иного человека. Нужен более формальный критерий.

Брингсйорд, Белло и Ферруччи учли эти замечания и формализовали идею Лавлейс следующим образом:

*Человек **H*** разработал *искусственный интеллект **A**, создающий рассказы. Если **A** создает рассказ **o**, который не является ошибкой, и **H** не может объяснить, как **o** был создан, то **A** следует рассматривать как способный к творчеству.*

Сделаем несколько уточнений:

1) **o** не обязательно должен быть рассказом, это может быть любое произведение или продукт;

2) в распоряжении **H** должна быть полная информация об архитектуре и принципах работы **A**. При необходимости **H** может изучить любой элемент **A**;

3) удачным объяснением работы **A** будет считаться воспроизведение результатов его деятельности с помощью простой манипуляции символами. Иными словами, **H** должен продемонстрировать четкий алгоритм действий;

4) поиск объяснения может занять сколько угодно времени, но в пределах разумного – не более нескольких лет.

В своей статье Брингсйорд, Белло и Ферруччи утверждают, что современные системы искусственного интеллекта не могут пройти тест Лавлейс. Компьютеры, которые пишут истории, ничего не создают. На самом деле разработчики потратили годы, чтобы выработать алгоритм создания историй. Несовершенства работы системы компенсируются добавлением все новых правил ad hoc. Это относится даже к обучающимся машинам.

Таким образом, тест Лавлейс предлагает альтернативный способ оценки способности искусственного интеллекта демонстрировать разумное поведение, неотличимое от человеческого, заменив способность поддерживать диалог на способность создавать рассказы.

К сожалению, в действительности тест Лавлейс является непроходимым. Любой **H**, имеющий время и необходимые ресурсы для создания **A**, сможет объяснить **o** [7]. Ведь наличие полной информации об архитектуре и принципах работы системы подразумевает понимание того, как она создает что-либо. Если же **o** не является результатом случайности, что оговорено в самих условиях теста, то необходимо должен существовать алгоритм его создания, пусть даже очень сложный.

Можно пойти дальше, задав следующий вопрос: а что если однажды будет обнаружен «алгоритм» человеческого творчества? Должны ли мы в этом случае признать человека не способным к творчеству? Независимо от ответа на вопрос тест Лавлейс оказывается бессмысленным. Либо никто не способен к творчеству, либо человек считается способным к творчеству априори.

Важно отметить и то, что человеку не требуется специальная теория творчества, чтобы творить. Обучение позволяет увеличить мастерство, сложность и качество произведений, но творчество как таковое доступно даже маленьким детям. У них есть знание-как без знания-что [1, с. 7]. Искусственный интеллект в то же время нуждается в наборе однозначных правил. Они могут быть предзаданными или выработанными в процессе «обучения» нейронной сети, но это всегда конкретный набор правил. При таких условиях любое действие искусственного интеллекта будет принципиально объяснимым, потому тест Лавлейс будет непроходимым априори.

Но допустим, что **H** не смог объяснить, как **A** создал **o**. Значит ли это, что мы должны признать, что **A** успешно прошел тест Лавлейс? Едва ли. Само условие, ограничивающее время, отведенное на объяснение, абсолютно произвольно. Почему **H** отводится всего несколько лет? Что если объяснение займет 10 или даже 100 лет? У нас нет оснований считать, что объяснение, поиск которого занял год, качественно отличается от объяснения, на которое пришлось затратить десятилетия. Это замечание тем более примечательно, если учесть, что ответы на многие вопросы наука получает лишь спустя столетия исследований, проводимых поколениями ученых.

Позднее Марк О. Ридл попытался переработать тест Лавлейс и создал его вторую версию – тест Лавлейс 2.0 [7]. Он формулируется следующим образом:

***A** создает артефакт **o** типа **t**. При этом **o** отвечает множеству условий **C**, где каждое условие $\boldsymbol{c_i} \in \boldsymbol{C}$ выразимо на естественном языке. **H** выбирает **t** и **C**, а затем оценивает, насколько созданный **o** им соответствует. Рефери **R** проверяет, чтобы сочетание **t** и **C** было возможным.*

*Способность **A** создать **o**, отвечающее требованиям **t** и **C**, является сильным индикатором интеллекта.*

Вторая версия теста Лавлейс частично решает проблемы первой версии. По крайней мере, тест становится проходимым, ведь есть рефери, который проверяет выполнимость заданных условий. Однако вторая версия теста все равно страдает от проблемы произвольности избранных критериев оценки. Не понятно ни то, почему нам следует доверять оценкам **H** и **R**, ни то, почему способность **A** создать **o** является надежным индикатором интеллекта.

Обе версии теста Лавлейс предполагают, что создание вымышленных историй требует ряда когнитивных способностей человеческого уровня. Проблема заключается в том, что в действительности мы этого не знаем, поскольку понятие «творчество» является достаточно проблемным.

Проблема понятия «творчество». Уже в античности творчество считалось отличительной чертой человека, признаком его разумности. Конечно, эта способность приписывалась также различным мифическим существам, однако все они обладали разумом человеческого типа. Поэтому данное уточнение не меняет сути, и мы не будем к нему возвращаться в дальнейшем.

Важно отметить, что до сих пор нет единой научной теории творчества. Сложно даже дать конкретное определение этому понятию [2, с. 7]. Однако если говорить в самых общих чертах, то творчество можно определить как деятельность по созданию чего-либо, которая характеризуется:

1) *оригинальностью* – творец может вдохновляться и опираться на другие произведения, однако он обязательно должен привносить что-то новое;

2) *необусловленностью* – предполагается, что человек творит не ради денег, славы и т. д., а в силу внутреннего желания;

3) *спонтанностью* – хотя выполнение творческой задачи может требовать особых навыков и соблюдения правил, их правильное применение само по себе не гарантирует достижения успеха.

Как отмечал Б. Гизелин, различие между творческой и нетворческой деятельностью является скорее субъективным [2, с. 12]. И действительно, ни одна из перечисленных характеристик творчества не может быть однозначно измерена. Остановимся на этом подробней.

Начнем с оригинальности. Строго говоря, не существует ничего полностью уникального. Художник ограничен видимым цветовым спектром, музыкант – слышимым диапазоном звука, писатель – выразительными способностями языка. Более того, идеи для своих произведений творец заимствует из окружающего мира, книг, других произведений искусства и т. д. На это указывал и сам Тьюринг в своем ответе леди Лавлейс [5, с. 79–85]. Таким образом, абсолютная оригинальность невозможна.

Однако возникает вопрос: какая степень оригинальности достаточна, чтобы назвать деятельность творческой? Можно ли это как-то измерить? Где проходит граница между еще не оригинальным и уже оригинальным? Эти и многие другие вопросы рождаются сами собой. Вероятно, на них можно ответить в рамках той или иной концепции творчества. Однако, как было отмечено выше, пока нет согласия по поводу того, какая именно теория творчества верна. Следовательно, мы не можем дать однозначный критерий оригинальности.

Что касается необусловленности, то мы вынуждены верить творцу на слово, когда он говорит, что творчество для него важнее всего. Биографические факты могут делать его утверждения о собственной мотивации более или менее правдоподобными, но окончательного доказательства мы не получим. Даже художник, живущий в нищете, может думать только о деньгах, просто ему не хватило таланта или удачи, чтобы продать свои картины. Возможна обратная ситуация: богатый художник всю жизнь творил только ради удовольствия, лишь случайно став успешным.

Также неверно было бы описывать мотивацию человека как нечто монолитное и неизменное. Например, мы легко можем согласиться, что художник может рисовать лишь ради творчества в юности, но с годами стать циничным, продолжая работать лишь ради денег. При этом мотивация человека часто является суммой сразу нескольких желаний и целей. Человек может искренне любить творчество, но не менее искренне желать денег и славы. Едва ли здесь возможно измерить процентное соотношение. Более того, даже для самого творца могут быть не до конца очевидны его мотивы.

Есть еще одна проблема, связанная с необусловленностью. Изменим ли мы отношение к произведению искусства, которое считаем великим, если узнаем, что оно было создано только ради денег? Если нет, то похоже, что мотивация художника все же не имеет значения. Если да, то нечто становится или перестает быть произведением искусства исключительно в силу общественного мнения о мотивации автора. Что примечательно, искусственный творец мог бы продемонстрировать большую необусловленность, чем любой из когда-либо живущих людей.

Наконец, спонтанность также является скорее проблемой для прояснения понятия «творчество». Сложно или даже невозможно обнаружить в этом процессе какие-либо методы и техники, гарантирующие успех. Кажется, что это соответствует интуитивному представлению о творчестве, но, что важнее, к этому мнению склоняются многие теоретики. Например, с точки зрения А. Л. Галина, творчество высшего типа необходимо требует использования интуиции [2, с. 22]. Я. А. Пономарев и вовсе утверждает, что сама идея «логики открытий» противоречит смыслу понятия «творчество» [2, с. 28]. Следовательно, способ создания произведения всегда сохраняет принципиальную непрояснённость и даже таинственность.

Итак, все три выделенные нами характеристики творчества делают его непригодным для использования в качестве критерия: оригинальность – слишком неоднозначный критерий, необусловленность в действительности не имеет существенного значения, а спонтанность делает понятие «творчество» принципиально непроясняемым.

Дополнительно можно задать следующий вопрос: все ли люди способны к творчеству? Ответ будет зависеть от того, насколько широко мы трактуем понятие «творчество». Сужение понятия «творчество» приведет нас к выводу, что не все люди разумны, например, если творчеством мы готовы назвать только создание произведений искусства, научных теорий и т. п. Расширение понятия, напротив, вынудит нас признать способность к творчеству и разумность за животными, машинами или даже, к примеру, погодными явлениями, если творчеством мы называем любое создание чего-то нового. Также мы рискуем начать решать задачу от ответа, подбирая такое определение понятию «творчество», которое соответствует нашим представлениям о способности людей к творчеству.

Таким образом, у нас нет однозначного мнения насчет способности людей к творчеству. Но в таком случае мы должны признать, что несправедливо требовать от машины доказывать разумность, демонстрируя способность к творчеству, если мы не можем достаточно однозначно определить это понятие. Велик риск того, что наша интерпретация понятия «творчество» будет напрямую зависеть от желания доказать или опровергнуть способность искусственного интеллекта к творчеству.

Обе версии теста Лавлейс страдают от описанной проблемы. Они предлагают произвольные критерии для оценки способности искусственного интеллекта к творчеству. Поэтому, с одной стороны, нет уверенности, что успешное прохождение теста доказывает наличие у агента способности к творчеству. Всегда есть вероятность, что проверка была недостаточно тщательной. С другой стороны, не является очевидным, что неспособность пройти тест Лавлейс доказывает неспособность агента к творчеству. Вполне возможно, что требование является принципиально невыполнимым. Следовательно, у нас нет оснований считать, что тест Лавлейс проверяет что-то большее, чем способность пройти тест Лавлейс.

Заключение. Как тест Тьюринга, так и обе версии теста Лавлейс по сути являются бихевиористскими тестами, т. е. они пытаются оценить наличие способности мыслить по внешним проявлениям. Однако здесь упускается сама суть вопроса о способности машины мыслить. В случае с человеком мы, как правило, не ставим такой вопрос, а изначально предполагаем за ним способность мыслить. Основанием обычно является примерно следующее рассуждение:

Я человек.

Я способен мыслить.

***X** тоже человек.*

*Следовательно, **X** тоже способен мыслить.*

Хотя это не является строгим доказательством, подобный довод имеет некоторую убедительность. Отсюда мы можем предположить, что за поведенческими реакциями других людей стоят те же мыслительные процессы, которые мы обнаруживаем у себя.

Но в случае с искусственным интеллектом у нас нет оснований для подобного предположения, ведь машина не только не является человеком, но даже не похожа на него. И дело вовсе не во внешнем сходстве, добиться которого можно сравнительно легко. Проблема заключается в том, что принцип работы искусственного интеллекта значительно отличается от принципов работы человеческого мозга. Также известно, что компьютер и мозг по-разному решают одни и те же задачи. Следовательно, нам достоверно известно, что сколь угодно сложный искусственный интеллект не мыслит в том же смысле, что и человек. По этой причине до тех пор, пока структурная организация искусственного интеллекта не станет подобной человеческому мозгу, бихевиористские тесты не будут иметь никакого значения.

Список использованных источников

1. Дрейфус, Х. Чего не могут вычислительные машины: критика искусственного разума / Х. Дрейфус ; пер. с англ. – Изд. 2-е. – М. : ЛИБРОКОМ, 2010. – 336 с.

2. Ильин, Е. П. Психология творчества, креативности, одаренности / Е. П. Ильин. – СПб. : Питер, 2008. – 433 с.

3. Оркестр сыграл законченную ИИ симфонию [Электронный ресурс]. – Режим доступа: https://naked-science.ru/article/hi-tech/orkestr-sygral-zakonchennuyu-ii. – Дата доступа: 21.02.2021.

4. Серл, Д. Сознание, мозг и программы / Д. Серл // Аналитическая философия: становление и развитие : антология : [пер. с англ., нем.] / общ. ред. и сост. А. Ф. Грязнова. – М., 1998. – С. 376–400.

5. Тьюринг, А. Вычислительные машины и разум / А. Тьюринг ; [пер. с англ. К. Королева]. – М. : АСТ, 2018. – 128 с.

6. Bringsjord, S. Creativity, the Turing test, and the (better) lovelace test / S. Bringsjord, P. Bello, D. A. Ferrucci // The Turing test: the elusive standard of artificial intelligence / ed. J. H. Moor. – Dordrecht, 2003. – P. 215–239. https://doi.org/10.1007/978-94-010-0105-2_12

7. Riedl, M. O. The Lovelace 2.0 Test of artificial creativity and intelligence [Electronic resource] / M. O. Riedl. – Mode of access: https://arxiv.org/pdf/1410.6142.pdf. – Date of access: 21.02.2021.

References

1. Dreyfus H. L. *What computers can't do: a critique of artificial reason*. Cambridge, MIT Press, 1992. 429 p.

2. Il'in E. P. *Psychology of creativity, creativity, giftedness.* St. Petersburg, Piter Publ., 2008. 433 p. (in Russian).

3. *The orchestra played a symphony completed by AI*. Available at: https://naked-science.ru/article/hi-tech/orkestr-sygral-zakonchennuyu-ii (accessed 21.02.2021) (in Russian).

4. Searle J. Minds, brains, and programs. *Behavioral and Brain Sciences,* 1980, vol. 3, no. 3, pp. 417–424. https://doi.org/10.1017/s0140525x00005756

5. Turing A. *Computing machinery and intelligence*. Moscow, AST Publ., 2018. 128 p. (in Russian).

6. Bringsjord S., Bello P., Ferrucci D. A. Creativity, the Turing Test, and the (Better) Lovelace Test. *The Turing test: the elusive standard of artificial intelligence*. Dordrecht, 2003, pp. 215–239. https://doi.org/10.1007/978-94-010-0105-2_12

7. Riedl M. O. *The Lovelace 2.0 Test of artificial creativity and intelligence*. Available at: https://arxiv.org/pdf/1410.6142.pdf (accessed 21.02.2021).

Информация об авторе

Ставровский Игорь Константинович – младший научный сотрудник. Институт философии, Национальная академии наук Беларуси (ул. Сурганова 1, корп. 2, 220072, Минск, Республика Беларусь). E-mail: tutoriks@gmail.com

Information about the author

Igor K. Stavrovsky – Junior Scientific Researcher. Institute of Philosophy of the National Academy of Sciences of Belarus (1 Surganov Str., Bldg 2, 220072 Minsk, Belarus). E-mail: tutoriks@gmail.com