ПРОГРЕСС РАЗРАБОТОК ИНТЕРФЕЙСА СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И ПРОБЛЕМА ИСКУССТВЕННОЙ ЛИЧНОСТИ

И. Ю. Ларионов
Санкт-Петербургский государственный университет
Санкт-Петербург

Автором выдвигается гипотеза, что одним из доминирующих направлений в области разработки искусственного интеллекта в XX в. было создание такой системы ИИ, которую пользователь не мог бы отличить от настоящей человеческой личности. Приведены и проанализированы примеры таких разработок. Встраивание элементов «искусственной личности» помогает решать ряд инженерных задач, однако из этого не следует, что данный ИИ может считаться искусственной личностью. Даны примеры публикаций в СМИ, в которых искусственный интеллект описывается как личность. Приведены аргументы в пользу того, что на этом основании нельзя переносить на ИИ и другие свойства личности.

Ключевые слова: искусственный интеллект, искусственная личность, нейросеть, чат-бот, интерфейс

PROGRESS OF DEVELOPMENT OF ARTIFICIAL INTELLIGENCE SYSTEMS INTERFACE AND PROBLEM OF ARTIFICIAL PERSONALITY

I. Yu. Larionov
Saint-Petersburg State University
Saint-Petersburg

The author puts forward the hypothesis that one of the dominant trends in the development of artificial intelligence in the XX century was the creation of such an AI system, which the user could not distinguish from a real human personality. Examples of such developments are given and analyzed. Embedding elements of "artificial personality" helps to solve a number of engineering problems, but it does not follow that this AI can be considered an artificial person. Examples are given of media publications in which artificial intelligence is described as a personality. Arguments are given that other properties of personality cannot be transferred to the AI on this basis.

Keywords: artificial intelligence, artificial personality, neural network, chatbot, interface

Взрывное распространения в начале 2023 года новостей об успехах разработок нейросетей следующего поколения вызвало множество публикаций СМИ, в которых эти системы описываются как некие самостоятельные сущности. Нередко высказываются опасения, что искусственный интеллект «оставит без работы», «заберёт работу» у целых групп профессий. Таким образом, проблематика искусственной агентности, возможности искусственной личности в общественном сознании остаётся актуальной.

Терминологически «искусственный интеллект» (artificial intelligence) не подразумевает искусственной личности, робота из классической научной фантастики и т. п. Однако он нередко описывается таким образом. Можно выделить два наиболее распространенных типа такого описания:

– ИИ представляется как самостоятельный субъект действия (напр. «Китайский ИИ уже создаёт проекты военных кораблей лучше человека» – https://www.vesti.ru/nauka/article/3246657; ср. та же новость в другом источнике: «В Китае протестировали нейросеть для разработки военных кораблей» – https://www.google.com/amp/s/m.gazeta.ru/amp/tech/news/2023/03/12/19947811.shtml);
– ИИ приписываются целенаправленные решения и действия (напр. «Искусственный интеллект: спасет или уничтожит человечество» – https://radiokp.ru/podcast/dialogi/679569; в самой статье подчеркивается, что нейросети самостоятельным интеллектом не обладают).

То, что пользователи не могут отличить результаты, выдаваемые нейросетью в определенном режиме ее работы, не значит, что можно автоматически переносить на нее и другие свойства личности. Цель доклада не в разворачивании мысли о несостоятельности подробного рода переноса способностей личности на современные системы искусственного интеллекта. Автором высказывается гипотеза, что сами разработки ИИ шли в направлении создания такой системы ИИ, которую пользователь был бы не в состоянии отличить от личности человека, и мы наблюдаем не более чем успешную (пусть и крайне успешную) реализацию такой технической задачи.

Ориентация на создание системы, которая будет в первую очередь восприниматься как человекоподобная, лежит в самом основании компьютерных разработок – знаменитый тест Тьюринга состоял в том, что нужно определить, взаимодействуете ли вы с искусственным интеллектом или с другим человеком. Описание данного теста было опубликовано Аланом Тьюрингом в философском журнале «Mind» [1].

Успешное прохождение той или иной программой данного теста стало одной из целей современных разработок в области искусственного интеллекта и нередко рассматривается как важный критерий оценки работы его создателей.

Важнейшее последствие популярности как статьи, так и самого теста А. Тьюринга в том, что под его влиянием закрепилась такая теоретическая конструкция, в которой наблюдаемые действия потенциального искусственного интеллекта сравниваются с проявлениями человеческого сознания. Достижением разработчиков искусственного интеллекта стало считаться такое поведение системы, которое неотличимо того или иного типа поведения существа, заведомо обладающего интеллектом естественным. Таким образом, речь идет о нахождении и определении критерия, по которому мы могли бы определить, что задача создания полноценного искусственного интеллекта выполнена успешно.

Сам А. Тьюринг открыто говорил в своей статье, что его интересует проблема, могут ли компьютеры именно мыслить, и его аргументы имеют силу, в первую очередь, для интеллектуалов, людей, для которых мышление является наибольшей ценностью и которые верят в превосходство человека как существа, обладающего данной способностью. Другие составляющие человеческой личности (и среди них – эмоции, переживания) не охватываются тестом Тьюринга.

Однако даже это не освободило предложение А. Тьюринга от справедливой критики.

Наиболее значимым критиком теста Тьюринга «на его собственном поле» проблематизации мышления, понимания, когнитивных способностей стал Дж. Сёрл. Для гипотетического состояния программы, при котором ее можно считать полноценно обладающей когнитивными состояниями, понимающей и т.п., Сёрл вводит термин «силь-ный искусственный интеллект» (Strong Artificial Intelligence). При этом, «только машины и могут мыслить, и в самом деле только очень особые виды машин, а именно мозги и машины, обладающие теми же каузальными способ-ностями, что и мозги. И это самое главное основание, почему силь-ный AI так мало рассказал нам о мышлении, ибо ему нечего сказать нам о машинах. По своему собственному определению, он касается программ, а программы – не суть машины» [2, с. 400]. Дж. Сёрл предложил мыслительный эксперимент под названием «китайская комната». Представим, что в комнате заперт человек, не владеющий китайским языком, но располагающий полным набором иероглифов и исчерпывающей инструкцией, как их соединять так, чтобы получился ответ на вопрос, заданный при помощи иероглифов, которые ему просовывают снаружи настоящие китайцы. Тогда у китайцев снаружи может возникнуть впечатление, что с ними общаются на их языке, хотя человек в комнате не понимает диалога, который он «ведет» и просто следует очень сложному алгоритму. Личность не просто пользуется словами по определенным правилам, реагируя на вопросы, и даже при определенных условиях становится инициатором разговора. Личность понимает слова и то, как именно ими пользуется. Тем самым, даже если мы в тесте Тьюринга перестанем отличать, человек с нами общается или «машина», из этого вовсе не будет следовать, что на другом конце коммуникации имеется личность в собственном смысле слова.

Среди современных критериев определения искусственного сознания посредством тестов большое влияние приобрели идеи Б. Герцеля (Ben Goertzel), известного разработчика и новатора в области разработки искусственного интеллекта. В короткой статье, опубликованной в научно-популярном журнале «New Scientist», он предлагает т.н. Robot College Student test: робот с искусственным интеллектом должен как студент пройти полноценный курс обучения и получить диплом. Если он выполнит это успешно, то с высокой вероятностью, по мнению Б. Герцеля, сознание и опыт такой машины можно признать похожими на человеческие, включая способности действовать в сложной обстановке, а также избирательный и творческий подход к получению и обработке информации [3].

Другим известным тестом, предложенным Б. Герцелем, является т.н. «кофейный тест» (The Coffee Test). Исходная идея принадлежит одному из основателей компании «Apple» Стивену Возняку (Steve Wozniak), неоднократно в своих выступлениях высказывавшего мысль, что, если робот сможет, войдя в любой дом, сориентироваться в незнакомом окружении и, самостоятельно найдя все необходимое, сварить чашку кофе, тогда можно говорить о наличии у него искусственного интеллекта. Сам С. Возняк сомневался в возможности создания подобного робота. Такое относительно простое действие, как приготовление чашки кофе, оказывается результатом комплекса сложных процессов в сознании человека. Тем самым, для того чтобы искусственный интеллект смог выполнить данную задачу успешно, требуется не просто большие мощности и разработанная программа, но и определенная «архитектура» различных способностей, среди которых разные виды памяти, способность к активному самообучению, использование ассоциаций и аналогий, мотивация и способность к спонтанному действию, коммуникация, направленная на развитие социальных отношений, а также самоконтроль и определенный образ себя [4].

Тем самым, критериями определения искусственного интеллекта стали, с одной стороны, способности решать сложные познавательные и творческие задачи, самообучаться, а с другой – имитировать социальное поведение человека.

Говоря о том, почему современные системы искусственного интеллекта не могут считаться искусственной личностью или полноценным искусственным агентом – в любом, даже самом ограниченном смысле («сильным искусственным интеллектом», «генерализованным (общим) искусственным интеллектом» и т. п.), из множества свойств, предлагаемых современной философией, хотелось бы выделить два критерия, понятные одновременно и философам, и разработчикам систем ИИ [5, 6]:

– автономия (имеющиеся системы нельзя признать автономными в полном смысле слова: удаленность от разработчика или оператора или эффективное функционирование без постоянного их вмешательства не могут быть критериями автономности);
– интенциональность (не как психологический феномен воли и/или намерения, но как процесс осознания и оценки ситуации с применением определенной ценностной парадигмы).

Иное направление в понимании агентности ИИ как раз и демонстрирует разработческий, инженерный подход к роли и назначениям ряда компонентов современной компьютерной системы. Встраивание элементов искусственной личности помогает решать ряд пользовательских задач. При этом в большинстве случаев мы имеем дело с определенным вариантом интерфейса программы. Одним из наиболее известных типов проектов искусственной личности являются голосовые помощники, отвечающие на вопросы пользователей и выполняющие команды. Они использует нейронные сети и алгоритмы машинного обучения для понимания запросов и поиска ответов. Интересным проектом является чат-бот Replika – приложение, создающее виртуального друга на основе данных пользователя. Приложение Woebot представляет собеседника-терапевта, который помогает людям бороться с депрессией и тревожностью. Не будем забывать о том, что персонифицированные образы уже очень давно используются в игровой индустрии в виде компьютерных персонажей, способных общаться с игроками, принимать решения и т.п.

Таким образом, в качестве свойств, воспринимаемых пользователем как элементы искусственной личности, можно назвать следующие признаки, проявляющие себя непосредственно во взаимодействии с пользователем:

– автономная работа (включая алгоритмы самообучения, требующая меньше непосредственно вводимых пользователем и/или разработчиком данных);
– эффективная коммуникация с человеком (в перспективе – ориентация в социальной реальности в широком смысле слова).

Исследование выполнено за счет гранта Российского научного фонда № 22-28-00379 «Трансформации морального агентства: этико-философский анализ» (https://rscf.ru/project/22-28-00379/).

ЛИТЕРАТУРА

1. Turing A. Computing Machinery and Intelligence // Mind. 1950, LIX (236). P. 433-460.
2. Сёрл Дж. Р. Сознание, мозг и программы // Аналитическая философия: становление и развитие. М.: Дом интеллектуальной книги, Прогресс-Традиция, 1998. С. 376-400.
3. Goertzel B. What counts as a conscious thinking machine? // New Scientist. 2012, September 5. URL: https://www.newscientist.com/article/mg21528813-600-what-counts-as-a-conscious-thinking-machine/ (дата обращения: 15.04.2023).
4. Goertzel B. Artificial General Intelligence: Concept, State of the Art, and Future Prospects // Journal of Artificial General Intelligence. 2014. Vol. 5. Iss. 1. P.1-46. DOI: 10.2478/jagi-2014-0001.
5. Anderson M., Anderson S.L. Machine Ethics: Creating an Ethical Intelligent Agent // AI Magazine. 2007. Vol. 28. Iss. 4. P.15-26. DOI: 10.1609/aimag.v28i4.2065.
6. Dennett D. When HAL Kills, Who's to Blame? Computer Ethics // HAL's Legacy: 2001's Computer as Dream and Reality / D. Stork (ed.). MIT Press, 1998.