УДК 17

И. Ю. Ларионов (*Санкт-Петербургский государственный университет*, e-mail: i.larionov@spbu.ru);
В. В. Головков (*Санкт-Петербургский государственный университет,* e-mail: golovkov.spb@mail.ru)

ПРОБЛЕМА ИМИТАЦИИ ЧЕЛОВЕЧЕСКОГО ИНТЕЛЛЕКТА И ОПИСАНИЕ РАЗРАБОТОК ИИ В СМИ

Исследование выполнено при финансовой поддержке РНФ № 24-28-00562 «Философские основания этических рисков в сфере систем искусственного интеллекта», Санкт-Петербургский государственный университет.

Рассматривается проблема искажения описания искусственного интеллекта в СМИ. Показывается глубинное основание восприятия ИИ через призму человеческого интеллекта. Описываются современные теоретические методы определения наличия интеллекта у машин. Практическая значимость результатов исследования в том, что искажение информации об ИИ в СМИ может способствовать формированию в обществе мнимых рисков использования ИИ. Обосновывается вывод, что одной из главных задач должна быть выработка научно обоснованной методологии выявления реальных угроз с учётом формирования СМИ представлений о развитии и функционировании ИИ.

Искусственный интеллект, риски, технологии, информация, рациональность, СМИ, сознание, тест Тьюринга

В современном мире вопросы, связанные с технологиями искусственного интеллекта, крайне актуальны. Связано это, во-первых, с экспоненциальным ростом развития технологии; во-вторых, с повсеместной интеграцией систем искусственного интеллекта в нашу повседневную жизнь; в-третьих, с созданным массовой культурой (посредством книг, фильмов, сериалов) восприятием искусственного интеллекта как чего-то непостижимого, удивительного, но в то же время, опасного. Именно эти три фактора на наш взгляд ключевым образом отражаются на информационном буме вокруг искусственного интеллекта, что выражается в бесконечных новостных заметках, статьях и видеороликах, паразитирующих на этой теме.

Гипотеза нашего исследования заключается в том, что современные СМИ продолжают тренд на восприятие искусственного интеллекта через призму интеллекта человеческого, возникший на заре появления ИИ.

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

1. ИИ представляется как самостоятельный субъект действия (например, «Китайский ИИ уже создаёт проекты военных кораблей лучше человека») [1] (для сравнения, та же новость в другом источнике: «В Китае протестировали нейросеть для разработки военных кораблей» [2]).

2. ИИ приписываются целенаправленные решения и действия (напр. «Искусственный интеллект: спасет или уничтожит человечество») [3]. В самой статье подчеркивается, что нейросети самостоятельным интеллектом не обладают.

Приведенные примеры можно многократно умножить.

Действительно, пользователи не всегда могут отличить результаты, выдаваемые современными нейросетями в определенном режиме их работы, с целесообразной активностью личности, но из этого вовсе не следует что можно автоматически переносить на нее все свойства личности. Однако с концептуальной точки зрения важнее учесть не общую необоснованность подобного рода переноса способностей личности на современные системы искусственного интеллекта. Важнее то, 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 крайне успешную реализацию данной технической задачи, но не более того.

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 [4]. Собственно, 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 от того или иного типа поведения существа, заведомо обладающего интеллектом естественным (например, знаменитые роботы Boston Dynamics, напоминающие по образу и «повадкам» собак, не говоря уже об антропоморфных машинах, таких как робот София). 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 «сильный 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 способностями, что и мозги. И это самое главное основание, почему сильный 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» [5, 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 [6].

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, требуются 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 [7].

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. Однако, здесь исследователи упираются, в том числе, в ограниченность знаний о нашем собственном интеллекте, не говоря уже о сознании. Проблемы возникают даже с определением уровня умственных способностей человека, которые упираются в однобокий и довольно конкретный тип мышления (как, например, знаменитый тест IQ). Как нам кажется, именно эта область требует глубоких философских исследований, которые позволили бы вывести более чёткие критерии определения наличия интеллекта у машин.

В то же время, речь идет именно о разработческой, инженерной концепции определенного компонента 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.

В заключение наше исследование приводит нас к выводу о том, что «горячие» заголовки новостных изданий продиктованы не только провокационными интенциями. Они имеют более глубинные основания, которые уходят к

самым ранним исследования искусственного интеллекта. Во многом, именно это основание задало тренд на соизмеримость создаваемых интеллектуальных машин с интеллектом человека, что и отражается как в массовой культуре, так и в СМИ. Актуальность нашего вывода и его практическая значимость в том, что в результате подобной «подачи» информации в обществе может сформироваться представление о многочисленных мнимых рисках использования ИИ. Разумеется, риски (притом вполне реальные) могут возникать как из-за неправильного использования прорывных технологий, так и в результате особенностей самого устройства и работы ИИ. Однако одной из насущных проблем будет выработка научно обоснованной методологии выявления и оценки подобных угроз с учетом фактора влияния ряда предрассудков на тех, кто несет ответственность за контроль и продвижение разработок и форм использования систем искусственного интеллекта.

Источники:

1. «Китайский ИИ уже создаёт проекты военных кораблей лучше человека» // Вести. URL: https://www.vesti.ru/nauka/article/3246657 (дата обращения: 16.01.2024).

2. «В Китае протестировали нейросеть для разработки военных кораблей» // Газета.ru. URL: https://www.google.com/amp/s/m.gazeta.ru/amp/tech/news/2023/-03/12/19947811.shtml (дата обращения: 16.01.2024).

3. «Искусственный интеллект: спасет или уничтожит человечество» // Радио комсомольская правда. URL: https://radiokp.ru/podcast/dialogi/679569 (дата обращения: 16.01.2024).

4. Turing A. Computing Machinery and Intelligence // Mind. 1950, LIX(236). P. 433–460.

5. Сёрл Дж. Р. Сознание, мозг и программы // Аналитическая философия: становление и развитие. М.: Дом интеллектуальной книги, Прогресс-Традиция, 1998. С. 376–400.

6. Goertzel B. What counts as a conscious thinking machine? // New Scientist. 2012, September 5.

7. Goertzel B. Artificial General Intelligence: Concept, State of the Art, and Future Prospects // Journal of Artificial General Intelligence. 2014, 35(1). P. 1–46.

I. Y. Larionov *(St. Petersburg state university);*
V. V. Golovkov *(St. Petersburg state university)*

THE PROBLEM OF IMITATING HUMAN INTELLIGENCE AND THE DESCRIPTIONS OF AI ENGINEERING IN THE MEDIA

The problem of distortion of the description of artificial intelligence in the media is considered. The deep basis of AI perception through the prism of human intelligence is shown. Modern theoretical methods of determining the presence of in-

telligence in machines are described. The practical significance of the research results is that the distortion of information about AI in the media can contribute to the formation of imaginary risks of using AI in society. One of the main future tasks is to develop a scientifically grounded methodology for identifying real threats, taking into account the prejudices formed by people who control and promote AI.

Artificial intelligence, risks, technologies, information, rationality, media, consciousness, Turing test