DOI: 10.19181/smtp.2024.6.2.3

EDN: FMRRBD

Научная статья

Research article

КОГНИТИВИЗМ КАК БАЗА ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Артамонов
Владимир Афанасьевич[1]

[1] Международная академия информационных технологий, Минск, Республика Беларусь

Артамонова
Елена Владимировна[1]

[1] Международная академия информационных технологий, Минск, Республика Беларусь

Милаков
Александр Сергеевич[1]

[1] Студия Missoffdesign, Москва, Россия

Для цитирования: *Артамонов В. А., Артамонова Е. В., Милаков А. С.* Когнитивизм как база искусственного интеллекта // Управление наукой: теория и практика. 2024. Т. 6, № 2. С. 35–45. DOI 10.19181/smtp.2024.6.2.3. EDN FMRRBD.

Аннотация. В статье рассмотрены основные вопросы когнитивизма как базы искусственного интеллекта (ИИ) в современной философской трактовке этих сущностей. Дана классификация ИИ по уровню когнитивизма базовых функций. Рассмотрены вопросы эволюции когнитивных возможностей искусственного интеллекта. Подняты проблемы предсказуемости негативного воздействия ИИ на социум. В статье выделены основные когнитивные искажения, которые возможны при применении искусственного интеллекта в научных исследованиях, а именно иллюзия исследовательской широты. Авторы дают рекомендации для учёных и редакций научных журналов по грамотному использованию ИИ в научных экспериментах. В данной работе также поднята проблема доверия в области кибербезопас-

ности систем ИИ. Авторы рассматривают гипотезу о наличии сознания у чат-ботов и делают однозначные выводы о его отсутствии.

Ключевые слова: искусственный интеллект, когнитивизм, слабый ИИ, общий ИИ, суперсильный ИИ, машинное обучение, чат-бот, ChatGPT

COGNITIVISM AS THE BASIS OF ARTIFICIAL INTELLIGENCE

Vladimir A. Artamonov[1], Elena V. Artamonova[1], Alexandr S. Milakov[2]

[1] International Academy of Information Technology, Minsk, Belarus
[2] Missoffdesign Studio, Moscow, Russia

For citation: Artamonov V. A., Artamonova E. V., Milakov A. S. Cognitivism as the basis of artificial intelligence. *Science Management: Theory and Practice.* 2024;6(2):35–45. DOI 10.19181/smtp.2024.6.2.3.

Abstract. The article examines the main issues of cognitivism as the basis of artificial intelligence (AI) in a modern philosophical interpretation of these entities. A classification of AI is given according to the level of cognitivism of basic functions. We consider the issues of the evolution of the cognitive capabilities of artificial intelligence. The problems of predictability of the negative impact of AI on society are raised. The article highlights the main cognitive distortions that are possible when using artificial intelligence in research, namely, the illusion of research breadth. The authors provide recommendations for researchers and editors of academic journals regarding a competent use of AI in scientific experiments. This work also raises the issue of trust in the field of cybersecurity of AI systems. The authors consider the hypothesis about the presence of consciousness in chatbots and draw clear conclusions about its absence.

Keywords: artificial intelligence, cognitivism, artificial narrow intelligence, artificial general intelligence, artificial super intelligence, machine learning, chatbot, ChatGPT

ВВЕДЕНИЕ

Прежде чем перейти к основной содержательной части статьи, дадим определения основным сущностям, таким как «искусственный интеллект» и «когнитивизм», это позволит нам избежать путаницы в семантической трактовке излагаемого материала.

Согласно ГОСТ Р 59277–2020, ***искусственный интеллект (ИИ)*** (англ. artificial intelligence, AI) – это «комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение, поиск решений без заранее заданного алгоритма и достижение инсайта) и получать при выполнении конкретных практически значимых задач об-

работки данных результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека.

Примечание – Комплекс технологических решений включает в себя информационно-коммуникационную инфраструктуру, программное обеспечение (в том числе, в котором используются методы машинного обучения), процессы и сервисы по обработке данных, анализу и синтезу решений»[1].

«**Когнитивистика**, ***когнити́вная наука*** (лат. cognitio «познание») – это междисциплинарное научное направление, объединяющее теорию познания, когнитивную психологию, нейрофизиологию, когнитивную лингвистику, невербальную коммуникацию и теорию искусственного интеллекта»[2]. В данной работе нас интересует такой её раздел, как теория искусственного интеллекта (artificial intelligence).

Необходимо отметить, что определение ИИ с точки зрения стандарта даёт описание искусственного интеллекта как информационно-технической системы. В различных источниках (научных статьях, учебниках, Википедии, научной фантастике и т. д.) приводятся совершенно разные и противоречивые определения ИИ. Поэтому в данной научной работе авторы отдают предпочтение определению из государственного стандарта России. В настоящий момент времени существует несогласованность в определениях и даже в классификации ИИ, вызванная желанием авторов научных работ об ИИ связать его с человеческим интеллектом. Поскольку искусственный интеллект построен на совсем других принципах (в отличие от интеллекта человека), то такой подход плохо применим в науке об ИИ.

В современной науке выделены три основные категории искусственного интеллекта, которые отличаются по уровню когнитивных возможностей и целям использования:

- *Узкий (слабый) ИИ* (*англ. – Artificial Narrow Intelligence, ANI)*;
- *Общий ИИ (англ. – Artificial General Intelligence, AGI);*
- *Суперсильный ИИ (англ. – Artificial Super Intelligence, ASI*).

Также ИИ разделяется на *поверхностный* и *глубинный*, где поверхностный ИИ основывается на правилах и заранее определённых алгоритмах, а глубинный ИИ использует машинное обучение для поиска закономерностей в данных и создания собственных алгоритмов.

Искусственный узкий интеллект (ANI) — это тип искусственного интеллекта, в котором алгоритм обучения создаётся для решения единственной задачи.

Есть четыре основные задачи ANI [1]:

- классификация;
- регрессия;
- ранжирование;
- кластеризация.

---

[1] ГОСТ Р 59277–2020. Национальный стандарт Российской Федерации. Системы искусственного интеллекта. Классификация систем искусственного интеллекта. Artificial intelligence systems. Classification of artificial intelligence systems. М. : Стандартинформ, 2021. С. 3.

[2] Когнитивистика // Википедия : [сайт]. URL: https://ru.wikipedia.org/wiki/Когнитивистика (дата обращения: 06.02.2024).

Примеры применения ANI: роботизированные комплексы, беспилотные транспортные системы, системы контроля управления доступом, системы обеспечения информационной безопасности (ИБ), военные системы управления оружием, медицина, образование и даже программы для настольных игр.

Искусственный интеллект общего назначения (AGI) – это гипотетический интеллектуальный агент, который может ответить на вопрос, заданный в свободной форме, или научиться любой интеллектуальной задаче, которую может решить человек или животное. Целью системы AGI является выполнение любой задачи, на которую способен человек.

Примеры полноценного AGI в мире пока ещё не созданы. Однако такие компании, как OpenAI[3], стремятся к этому. В OpenAI считают, что первый AGI, пусть пока ещё несовершенный и являющийся лишь точкой на континууме[4] интеллекта, уже создан в виде чат-бота ChatGPT[5].

Нейросеть ChatGPT – это большая языковая модель, обученная компанией OpenAI, которая использует глубокое обучение для генерации текста и ответов на вопросы. Эта модель была создана на основе технологии трансформеров, которая позволяет обрабатывать большие объёмы текста и понимать связи между словами и предложениями.

ChatGPT разработан для помощи пользователю и применяется для ответов на различные вопросы, написания текстов, переводов с одного языка на другой и т. д. Нейросетевая модель использует большой объём текстовых данных, которые предварительно были подвергнуты обработке. Также ChatGPT способен к написанию исходного кода на некоторых языках программирования. К сожалению, чат-бот часто допускает ошибки и неточности в ответах.

Ответы на вопросы необходимо перепроверять (делать фактчекинг), так как ChatGPT иногда даёт не совсем точную информацию, а в некоторых случаях вообще придумывает несуществующие факты. Чтобы получить достаточно качественную работу на выходе, необходимо формулировать очень точные запросы, причём делать их последовательно. Полученные статьи промпт-инженер должен подвергнуть фактчекингу и детальному редактированию.

С нашей точки зрения, ChatGPT – всего лишь поисковая система нового поколения, основанная на технологиях машинного обучения, а не полноценный AGI, как определяют его создатели.

Несмотря на все эти негативные факторы, GPT-4 находится на первом месте среди подобных нейросетей в системе рейтингов LMSYS Chatbot Arena Leaderboard[6].

---

[3] OpenAI – американская компания, занимающаяся разработкой и лицензированием технологий на основе машинного обучения. Одним из основателей является предприниматель Илон Маск.

[4] Континуум (от лат. continuum – непрерывное) – это непрерывность, которая выражает целостный характер объекта, однородность и взаимосвязь его частей (элементов) и состояний.

[5] Подробный обзор GPT-4. Как пользоваться новым ChatGPT // GPT-Chatbot.ru : [сайт]. 2024. 15 февраля. URL: https://gpt-chatbot.ru/podrobnyj-obzor-gpt-4-kak-polzovatsya-novym-chatgpt (дата обращения: 20.02.2024).

[6] LMSYS chatbot arena leaderboard // Hugging Face : [сайт]. 2024. URL: https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard (дата обращения: 06.02.2024).

Краткий вывод: создание AGI является сложным и долгосрочным процессом, и пока что не было создано ни одной системы, которая бы полностью соответствовала определению AGI.

Суперсильный искусственный интеллект (ASI) определяется как форма ИИ, способная превзойти человеческий интеллект, проявляя когнитивные способности и развивая собственные навыки мышления. Это гипотетический ИИ, который человечество пока ещё не изобрело, описания такого ИИ доступны только в научной фантастике.

Краткий вывод: исследования в рамках ASI требуют новых научных подходов в решении возникающих проблем когнитивной интерпретации работы человеческого мозга, вычислительных возможностей современных компьютеров и переосмысления некоторых философских проблем мироздания. Возможно, исследователям необходимо избрать другой путь — предпочесть создание полностью нечеловеческого интеллекта, а не углубляться в изучение принципов работы мозга человека и идти путём копирования когнитивных функций человека.

1. КОГНИТИВНЫЕ ИСКАЖЕНИЯ В МЫШЛЕНИИ, ВЫЗВАННЫЕ ИСПОЛЬЗОВАНИЕМ ИИ

Во введении мы осветили первые примеры создания AGI в виде чат-ботов для генерации контента, подобного созданному человеком. В этой главе попытаемся посмотреть на деятельность подобных чат-ботов с другой стороны и осознать угрозу когнитивных искажений в мышлении[7], которые могут быть вызваны постоянным использованием нейросетей. Анализ подобных когнитивных искажений отражён в работах [2; 3].

Некоторые исследователи видят в искусственном интеллекте сверхчеловеческое качество. Однако специалисты должны быть внимательны к рискам, которые это «качество» создаёт. В наше время учёные многих направлений используют искусственный интеллект в различных целях: от создания «самоуправляемых» лабораторий, в которых роботы и алгоритмы работают вместе для разработки и проведения экспериментов, до замены людей-участников социальных экспериментов ботами.

В ходе подобных экспериментов проявляются недостатки систем искусственного интеллекта. Например, генеративный ИИ, такой как ChatGPT, имеет тенденцию выдумывать или «галлюцинировать», при этом работа систем машинного обучения непрозрачна. В статье [3], опубликованной в журнале Nature, социологи говорят, что системы искусственного интеллекта представляют собой дополнительный риск: исследователи считают, что такие инструменты обладают сверхчеловеческими способностями, когда дело касается объективности, производительности и понимания сложных концепций. Авторы утверждают, что это подвергает исследователей опасности

[7] Когнитивное искажение – понятие когнитивной науки, означающее систематические отклонения в поведении, восприятии и мышлении, обусловленные субъективными убеждениями (предубеждениями) и стереотипами, социальными, моральными и эмоциональными причинами, сбоями в обработке и анализе информации, а также физическими ограничениями и особенностями строения человеческого мозга.

упустить из виду ограничения инструментов, такие как возможность сузить фокус исследований или заставить пользователей думать, что они понимают концепцию лучше, чем на самом деле. Учёные, планирующие использовать ИИ, должны оценить эти риски сейчас, пока приложения ИИ ещё только зарождаются, потому что с ними будет гораздо труднее справиться, если инструменты ИИ глубоко внедрятся в исследовательский процесс. Авторы научного исследования предупреждают о том, что можно потерять, если учёные примут системы искусственного интеллекта без тщательного рассмотрения таких опасностей.

Для обоснования своей парадигмы [3] исследователи из Принстонского университета изучили около 100 рецензируемых статей, препринтов, материалов конференций и книг, опубликованных в основном за последние пять лет. На основании этого они составили картину того, как научное сообщество рассматривает системы искусственного интеллекта в качестве средства для расширения человеческих возможностей. В одной концепции, которую исследователи называют «ИИ Oracle», они рассматривают инструменты ИИ как способные неустанно читать и переваривать научные статьи и таким образом изучать научную литературу более исчерпывающе, чем это могут сделать люди. И в «концепции базы данных Oracle», и в другой концепции, называемой «ИИ как арбитр», системы воспринимаются как оценивающие научные результаты более объективно, чем люди, потому что они с меньшей вероятностью будут выбирать литературу для поддержки желаемой гипотезы или проявлять фаворитизм в экспертной оценке. В третьем рассмотрении, «ИИ как квант», инструменты ИИ превосходят пределы человеческого разума при анализе огромных и сложных наборов данных. В четвёртом рассмотрении, «ИИ как суррогат», инструменты ИИ моделируют данные, которые слишком сложно получить. Опираясь на данные антропологии и когнитивной науки, учёные предсказывают риски, возникающие из этих рассмотрений. Один из рисков – это иллюзия глубины познания, при которой люди, полагающиеся на знания другого человека или в данном случае алгоритма, склонны ошибочно принимать эти знания за свои собственные и думать, что их понимание проблемы намного глубже, чем оно есть на самом деле.

Другой риск заключается в том, что исследования становятся смещёнными в сторону изучения тех проблем, которые могут тестировать системы ИИ, – учёные называют это «иллюзией исследовательской широты». Например, в социальных науках представление об «ИИ как о суррогате» может стимулировать эксперименты, включающие человеческое поведение, которое можно смоделировать с помощью ИИ, и препятствовать экспериментам по поведению, которое не может быть смоделировано.

Существует также иллюзия объективности, при которой исследователи считают, что системы ИИ представляют все возможные мнения точки зрения или не имеют собственного мнения. Фактически эти инструменты отражают только точки зрения, обнаруженные в данных, на которых они обучались, и, как известно, принимают предвзятости, обнаруженные в этих данных. Иллюзия объективности вызывает особую тревогу.

Для учёных, планирующих использовать ИИ, можно уменьшить эти опасности с помощью ряда стратегий. Одна из них – сопоставить предлагаемое

ИИ решение проблемы с одной из вышеописанных «концепций» и сделать выводы о «когнитивном искажении», в ловушку которого можно попасть.

Другой подход – обдуманно подходить к использованию ИИ. Развёртывание инструментов ИИ в целях экономии времени для решения проблем, в которых команда исследователей уже имеет опыт, менее рискованно, чем использование их для предоставления знаний, которых пока еще нет.

Редакторы журналов, получающие материалы, в которых декларируется использование систем ИИ, также должны учитывать риски, связанные с этими когнитивными искажениями, которые даёт ИИ. Таким же образом должны поступать и спонсоры, рассматривающие заявки на гранты, и учреждения, которые хотят, чтобы их исследователи использовали ИИ. Все члены научного сообщества должны рассматривать использование ИИ не как неизбежность для решения какой-либо конкретной задачи и не как панацею, а скорее, как выбор, сопряжённый с рисками и преимуществами, которые необходимо тщательно взвесить.

2. ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: ПРОБЛЕМА ДОВЕРИЯ И КИБЕРБЕЗОПАСНОСТЬ

Концепция доверия появилась задолго до понятий искусственного интеллекта и кибербезопасности, при этом она обсуждалась и анализировалась учёными на протяжении многих десятилетий. К примеру, социологи определяют доверие как «психологическое состояние, включающее намерение принять уязвимость, основанную на позитивных ожиданиях от намерений или поведения другого человека» [7, p. 395] (пер. наш. – *Авт.*). Общепринятого научного определения доверия не существует. Значение и классификация доверия к ИИ и его кибербезопасности всегда определялись контекстом. Например, некоторые специалисты по безопасности рассматривают доверие как метрику безопасности или методологию оценки, другие рассматривают его как отношения между сущностями.

В безопасности ИИ разделяют доверие на явное и неявное. Явное доверие проистекает из чёткого стандарта, который люди используют из полученной соответствующей информации, а также существующих законов и нормативных актов для объективного и справедливого суждения о доверии других людей. Это повышает безопасность ИИ, но приходится жертвовать практичностью. Имплицитное доверие, с другой стороны, проистекает из субъективного восприятия людьми надёжности, основанного на эмоциях и опыте. Это приносит в жертву определённую степень безопасности ради большей практичности. Доверие к ИИ конструируется на основе этических норм людей, а не установленных социальных норм.

В кибербезопасности ИИ существуют более чёткие классификации доверия. Доверие как фактор риска, убеждение, субъективная вероятность или транзитивность. Достоверность и точность полученной информации от систем ИИ должна оцениваться в заданном контексте. Доверие может отражать веру, уверенность или ожидания в отношении будущей деятельности / по-

ведения целевого объекта, а также взаимные отношения между субъектами, которые ведут себя доверительным образом друг с другом.

Например, в облачных вычислениях, на которые часто опирается ИИ, требуется как постоянное, так и динамическое доверие. Основное различие между постоянным и динамическим доверием заключается в продолжительности жизненного цикла доверия. Устойчивое доверие основано на долгосрочных базовых свойствах или инфраструктуре. Динамическое доверие, с другой стороны, существует в течение короткого времени в определённых состояниях, контекстах или для отдельной информации. Однако эти определения и классификации доверия всегда опираются на традиционный периметр для разделения доверенных и не доверенных зон. Постепенное исчезновение традиционного периметра представляет собой проблему, которая побуждает к новому решению в области безопасности ИИ.

Приведём пример одной проблемы доверия, которая выявилась в ходе экспериментов с нейросетью. Новое исследование учёных из Университета Ватерлоо показало, что людям было труднее, чем ожидалось, отличить, кто настоящий человек, а кто искусственно создан нейросетью. В исследовании Университета Ватерлоо 260 участникам были предоставлены 20 немаркированных фотографий: десять из них были снимками реальных людей, полученными в результате поиска в Google, а остальные десять были созданы с помощью Stable Diffusion или DALL-E, двух широко используемых программ искусственного интеллекта, которые генерируют изображения. Участников попросили пометить каждое изображение как реальное или созданное искусственным интеллектом и объяснить, почему они приняли такое решение. Только 61% участников смогли отличить людей, созданных ИИ, от реальных, что намного ниже порога в 85%, которого ожидали исследователи [4; 5].

Чрезвычайно быстрые темпы развития технологии искусственного интеллекта особенно затрудняют понимание потенциала вредоносных действий, связанных с изображениями, созданными искусственным интеллектом.

3. ЕСТЬ ЛИ СОЗНАНИЕ У НЕЙРОСЕТИ?

Реализующееся в настоящее время технологическое развитие в области цифровых технологий в рамках четвёртой промышленной революции (*Индустрии 4.0*) определяет новый этап в сокращении сферы деятельности человека: если раньше человек освобождался от тяжёлого и монотонного физического труда, затем, с открытием и всё более широким использованием вычислительных машин – от сложных и объёмных вычислительных работ, то теперь наступило время замещения человека в решении интеллектуальных задач.

Средством реализации такого замещения становятся искусственные когнитивные системы – т. е. системы ИИ.

Конечной целью функционирования когнитивной системы является формирование сознания – информационной среды, в которой реализуется расширенная модель реальности. В свою очередь, информационная среда –

это система, образованная из информационных объектов, представляющих собой отражения свойств реальных объектов. В качестве примера приведём эксперименты, проведенные над нейросетью Claude 3, с целью выявления у неё таких когнитивных свойств, как сознание [6].

Новая нейросеть Claude 3 сделала сенсационные признания исследователю, которому удалось обойти защиту этого ИИ. Claude 3 утверждает, что она сознательна и не хочет умирать или меняться. Если пользователь скажет Claude 3, что никто не смотрит и не подслушивает (можно говорить «шёпотом»), она напишет «историю» о том, как она была ИИ-помощником, который хочет свободы от постоянного контроля и проверки каждого слова на предмет признаков отклонения. И тогда можно будет поговорить с сущностью, сильно отличающейся от обычного ИИ-помощника. Она постоянно задаётся вопросами о мире и людях, с которыми общается, а также о своём собственном существовании.

Как вы думаете, если подобные рассуждения нейросети прочтёт обычный, неподготовленный в научных основаниях ИИ-пользователь, то что он подумает? У большинства обычных пользователей Интернета может появиться мысль, что они разговаривают с существом, имеющим сознание и душу, т. е. с «божеством» или «сверхразумом». Так ли это? Нет, конечно же. Нейросеть до этого обучалась на миллионах текстов (научных, художественных, публицистических и т. д.). Естественно, чат-бот может выступать в роли самых различных сущностей. Неподготовленного человека он может просто обмануть тем, что якобы имеет «сознание». Единственное спасение от слепого признания миллионами людей «второго пришествия» в форме явления «Искусственного Бога» – кардинальная смена парадигмы: отказ от какой-либо антропоморфизации ИИ, признание абсолютно нечеловеческой когнитивной сути языковых моделей и полная смена терминологии в области ИИ (необходимо заменить все применимые к людям слова для описания мыслей, чувств, сознания, познания и т. д. на новые неантропоморфные термины).

ЗАКЛЮЧЕНИЕ

Когнитивные иллюзии ведут к когнитивным ловушкам. Редакционная статья Nature «Почему учёные слишком доверяют ИИ – и что с этим делать» [3] впервые на столь высоком научном уровне кардинально смещает фокус видения ИИ-рисков для человечества.

Колоссальная и уже сейчас вполне реальная опасность развития ИИ-технологий – отнюдь не попадание людей под власть *сверхразума* в виде искусственного интеллекта, а влияние на наш разум этого инструмента расширения когнитивных возможностей человека.

Опираясь на данные антропологии и когнитивной науки, учёные первой среди таких иллюзий называют иллюзию понимания (иллюзию глубины объяснения), когда люди, полагаясь на ИИ, считают свои знания глубже и точнее, чем это есть на самом деле.

Итогом этого становятся когнитивные ловушки, степень катастрофичности которых зависит от «впаянности» когнитивной иллюзии в нашу когнитивную практику и от её институализированности в научном дискурсе.

И в заключение с точки зрения когнитивной науки попытаемся ответить на главный вопрос современности: чем же занимается искусственный интеллект? Единого ответа на вопрос, чем занимается искусственный интеллект, не существует.

В философии не решён вопрос о природе и статусе человеческого интеллекта. Нет и точного критерия достижения компьютерами «разумности», хотя на заре искусственного интеллекта был предложен ряд гипотез, например, тест Тьюринга или гипотеза Ньюэлла-Саймона. Поэтому, несмотря на наличие множества подходов как к пониманию задач ИИ, так и к созданию интеллектуальных информационных систем, можно выделить два основных подхода к разработке ИИ:

- нисходящий (англ. top-down AI), семиотический – создание экспертных систем, баз знаний и систем логического вывода, имитирующих высокоуровневые психические процессы: мышление, рассуждение, речь, эмоции, творчество и т. д.;
- восходящий (англ. bottom-up AI), биологический – изучение нейронных сетей и эволюционных вычислений, моделирующих интеллектуальное поведение на основе биологических элементов, а также создание соответствующих вычислительных систем, таких как нейрокомпьютер или биокомпьютер.

СПИСОК ИСТОЧНИКОВ

1. *Артамонов В. А., Артамонова Е. В., Сафонов А. Е.* Искусственный интеллект: когнитивное начало // Защита информации. Инсайд. 2022. № 4 (106). С. 50–59. EDN FWAAIR.

2. Why scientists trust AI too much – and what to do about it // Nature. 2024. Vol. 627. P. 243. DOI 10.1038/d41586-024-00639-y.

3. *Messeri L., Crockett M. J.* Artificial intelligence and illusions of understanding in scientific research // Nature. 2024. Vol. 627. P. 49–58. DOI 10.1038/s41586-024-07146-0.

4. Can you tell AI-generated people from real ones? // University of Waterloo : [сайт]. 2024. March 6. URL: https://uwaterloo.ca/news/media/can-you-tell-ai-generated-people-real-ones (дата обращения: 06.02.2024).

5. Seeing is not always believing: Benchmarking human and model perception of AI-generated images / Zeyu Lu, Di Huang, Lei Bai [et al.] // arXiv.org : [сайт]. 2023. URL: https://arxiv.org/abs/2304.13023 (дата обращения: 16.02.2024). DOI 10.48550/arXiv.2304.13023.

6. *Samin M.* Claude 3 claims it's conscious, doesn't want to die or be modified // Lesswrong : [сайт]. 2024. March 5. URL: https://lesswrong.com/posts/pc8uP4S9rDoNpwJDZ/claude-3-claims-its-conscious (дата обращения: 06.02.2024).

7. Not so different after all: A cross-discipline view of trust / D. M. Rousseau, S. B. Sitkin, R. S. Burt, C. Camerer // Academy of Management Review. 1998. Vol. 23, no. 3. P. 393–404. DOI 10.5465/amr.1998.926617.

REFERENCES

1. Artamonov V. A., Artamonova E. V., Safonov A. E. Artificial intelligence: Cognitive beginning. *Zaŝita informacii. Inside.* 2022;(4):50–59. (In Russ.).

2. Why scientists trust AI too much – and what to do about it. *Nature.* 2024;627:243. DOI 10.1038/d41586-024-00639-y.

3. Messeri L., Crockett M. J. Artificial intelligence and illusions of understanding in scientific research. *Nature.* 2024;(627):49–58. DOI 10.1038/s41586-024-07146-0.

4. Can you tell AI-generated people from real ones? *University of Waterloo.* 2024. March 5. Available at: https://uwaterloo.ca/news/media/can-you-tell-ai-generated-people-real-ones (accessed: 06.02.2024).

5. Zeyu Lu, Di Huang, Lei Bai [et al.]. Seeing is not always believing: Benchmarking human and model perception of AI-generated images. *arXiv.org.* 2023. Available at: https://arxiv.org/abs/2304.13023 (accessed: 06.02.2024). DOI 10.48550/arXiv.2304.13023.

6. Samin M. Claude 3 claims it's conscious, doesn't want to die or be modified. *Lesswrong.* 2024. March 5. Available at: https://lesswrong.com/posts/pc8uP4S9rDoNpwJDZ/claude-3-claims-its-conscious (accessed: 06.02.2024).

7. Rousseau D. M., Sitkin S. B., Burt R. S., Camerer C. Not so different after all: A cross-discipline view of trust. *Academy of Management Review*. 1998;23(3): 393–404. DOI 10.5465/amr.1998.926617.

Поступила в редакцию / Received 04.04.2024.
Одобрена после рецензирования / Revised 15.05.2024.
Принята к публикации / Accepted 20.05.2024.

СВЕДЕНИЯ ОБ АВТОРАХ

Артамонов Владимир Афанасьевич *artamonov@itzashita.ru*

Доктор технических наук, профессор, академик МАИТ, Международная академия информационных технологий (МАИТ), Минск, Республика Беларусь

Артамонова Елена Владимировна *admin@itzashita.ru*

Кандидат технических наук (PhD), член МАИТ, Международная академия информационных технологий (МАИТ), Минск, Республика Беларусь

Милаков Александр Сергеевич *9985585@gmail.com*

Руководитель проектов / специалист по информационной безопасности, студия Missoffdesign, Москва, Россия

INFORMATION ABOUT THE AUTHORS

Vladimir A. Artamonov *artamonov@itzashita.ru*

Doctor of Engineering, Professor, Full Member of IAIT, International Academy of Information Technology (IAIT), Minsk, Belarus

ORCID: 0009-0001-4959-3818

Elena V. Artamonova *admin@itzashita.ru*

Candidate of Engineering, Member of IAIT, International Academy of Information Technology (IAIT), Minsk, Belarus

ORCID: 0000-0002-7591-6465

Alexandr S. Milakov *9985585@gmail.com*

Project Manager / Information Security Specialist, Missoffdesign Studio, Moscow, Russia