ДИСКУРС. 2025. Т. 11, № 1. С. 16–30
DISCOURSE. 2025, vol. 11, no. 1, pp. 16–30

Оригинальная статья
УДК 172:004.8
http://doi.org/10.32603/2412-8562-2025-11-1-16-30

Новации современного искусственного интеллекта: ценностно-ориентированный подход

***Раиса Ильинична Мамина*[1]✉*, Анна Валерьевна Ильина*[2]**

[1, 2]*Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина), Санкт-Петербург, Россия*

[1]✉*maminaraisa@yandex.ru, https://orcid.org/0000-0003-3301-636X*
[2]*a.ilyina2045@gmail.com, https://orcid.org/0000-0003-3002-1841*

Введение. За последние годы в жизни современного социума роль искусственного интеллекта (ИИ) возросла кратно. Новации в области ИИ создают для человека и общества в целом не только новые возможности, но и риски, проблемы, угрозы, что приводит к актуализации значения риск-ориентированного подхода к регулированию создания и развития ИИ как на международном, так и на национальном уровнях. При этом проблема управления рисками имеет несколько взаимосвязанных направлений, одним из центральных считается проблема ценностных ориентиров, на которых выстраивается этика и этичность ИИ. В статье рассматриваются вопросы, связанные с ценностно-ориентированным направлением развития современного ИИ.

Методология и источники. Используется методология культур-философского, аксиологического и междисциплинарного подходов. В качестве источников в статье использованы научные исследования отечественных и зарубежных авторов, документы, публикации и сайты, посвященные современному состоянию ИИ и его проблематике.

Результаты и обсуждение. Тематика современного ИИ развивается как специальная область научного и дисциплинарного знания, а также как масштабируемая ИИ-индустрия. Один из последних трендов в рамках текущего ИИ – эмоциональный ИИ и его возможности в налаживании эффективной коммуникации с человеком. Однако несмотря на революционность новой ИИ-технологии и особую значимость этой новации в коммуникации ИИ с человеком, проблемы этики и ценностно-ориентированное развитие современного ИИ определяются специалистами как ключевые проблемы современности.

Заключение. Современный ИИ рассматривается сегодня с позиций различных классификаций, где центральное место занимает классификация, основанная на сопоставлении интеллекта человека и ИИ. Соответственно, формализация нравственных ценностей и этических принципов в процессе разработки и функционирования алгоритмов ИИ имеет важное значение для ценностно-ориентированного взаимодействия «человек – ИИ». Однако система общечеловеческих ценностей и ценностей ИИ совпадают лишь частично. Как следствие – актуализация новых подходов, не сопоставляющих ИИ и интеллект человека, в частности, на сегодняшний день междисциплинарный подход «4E Cognition» рассматривается специалистами как наиболее продуктивный во всех отношениях.

ДИСКУРС. 2025. Т. 11, № 1. С. 16–30
DISCOURSE. 2025, vol. 11, no. 1, pp. 16–30

Ключевые слова: цифровые технологии, прикладной ИИ, общий ИИ, этические принципы, общечеловеческие ценности, ценности ИИ, подход «4E Cognition»

Для цитирования: Мамина Р. И., Ильина А. В. Новации современного искусственного интеллекта: ценностно-ориентированный подход // ДИСКУРС. 2025. Т. 11, № 1. С. 16–30. DOI: 10.32603/2412-8562-2025-11-1-16-30.

Original paper

Innovations of Contemporary Artificial Intelligence: Value-Based Approach

***Raisa I. Mamina*[1]✉*, Anna V. Ilina*[2]**

[1, 2]*Saint Petersburg Electrotechnical University, St Petersburg, Russia*

[1]✉*maminaraisa@yandex.ru, https://orcid.org/0000-0003-3301-636X*

[2]*a.ilyina2045@gmail.com, https://orcid.org/0000-0003-3002-1841*

Introduction. In recent years, the role of artificial intelligence (AI) in social life has increased. Innovations in the field of AI create not only new opportunities for a person and society as a whole, but also risks, problems and threats, which leads to the actualization of a risk-oriented approach towards the regulation of AI development on both the international and national levels. The problem of AI risk management has several interrelated areas, one of the central ones is the problem of identification of values on which the ethics of AI is built. The article considers issues related to the value-oriented direction of development of contemporary AI.

Methodology and sources. In the article there were used the methodology of cultural-philosophical, axiological and interdisciplinary approaches is used. The sources used in the article are scientific research of domestic and foreign authors, documents, publications and websites devoted to the current state of AI and its problems.

Results and discussion. The topic of contemporary AI is developing as a special area of scientific and disciplinary knowledge, as well as a scalable AI industry. One of the latest trends in the current AI is emotional AI and its capabilities in establishing effective communication with person. However, despite the revolutionary nature of the new AI technology and the special significance of this innovation in AI communication with person, the problems of ethics and the value-oriented development of contemporary AI are defined by experts as key problems of our time.

Conclusion. Contemporary AI is analyzed today from the standpoint of various classifications, the central place is occupied by the classification based on the comparison of human intelligence and AI. In this regard, the formalization of moral values and ethical principles in the process of developing and operating AI algorithms is important for the value-oriented human-AI interaction. However, the system of universal human values and AI values coincide only partially. As a result, new approaches emerge that do not compare AI and human intelligence, in particular, the interdisciplinary approach «4E Cognition», which is considered by experts as the most productive in all respects.

Keywords: digital technologies, applied AI, general AI, ethical principles, universal values, AI values, “4E Cognition” approach

For citation: Mamina, R.I. and Ilina, A.V. (2025), “Innovations of Contemporary Artificial Intelligence: Value-Based Approach”, *DISCOURSE*, vol. 11, no. 1, pp. 16–30. DOI: 10.32603/2412-8562-2025-11-1-16-30 (Russia).

ДИСКУРС. 2025. Т. 11, № 1. С. 16–30
DISCOURSE. 2025, vol. 11, no. 1, pp. 16–30

Введение. В коммуникации «человек – ИИ» проблематика «moral machines» остается в настоящее время все еще нерешенной. При этом главным препятствием является то, что этичность ИИ имеет не ценностный и смысловой, а исключительно формализованный характер. Необходимость перевода базовых этических принципов в нравственные стандарты, усваиваемые машинами, как и проблема этичности ИИ в целом, стали сегодня предметом особого внимания международного научно-исследовательского поиска [1]. Перефразируя знаменитое изречение отца-основателя современного менеджмента Питера Друкера: «Культура съедает стратегию на завтрак» («Culture eats strategy for breakfast») [2], в котором подчеркнуто значение поведенческой культуры людей в организации и не только, для достижения намеченных целей применительно к ИИ можно сказать, что неэтичность ИИ может съесть на завтрак и культуру, и стратегию, и все цивилизационное развитие современного социума.

Методология и источники. Представленные в работе выводы опираются на методологию культур-философского, аксиологического и междисциплинарного подходов. Выполненный в статье анализ основывается на специальной научной литературе, западных и отечественных изданиях [3–5], а также Интернет-источниках и сайтах, которые посвящены современному состоянию ИИ и задачам, которые стоят перед специалистами в области ИИ и, в частности, этики ИИ.

Результаты и обсуждение. Современный ИИ, основанный на машинном обучении, специалисты характеризуют как системы, расширяющие потенциал человека благодаря их способности к обучению, распознаванию и осмыслению [4]. Для систематизации существующих разработок научное ИИ-сообщество представило ряд обобщающих классификаций, которые помогают упорядочить наличные представления о предметном поле исследовательских ИИ-практик. Одной из наиболее популярных на сегодняшний день является общепринятая в международном ИИ-сообществе классификация, которая строится на отличии видов ИИ по уровню их интеллектуальных возможностей в сравнении с интеллектом человека (ЕИ), целям использования и перспективам будущего развития. В рамках этой классификации сегодня выделяют прикладной ИИ (Narrow Artificial Intelligence/ANI/AI), общий ИИ (Artificial General Intelligence/AGI) и искусственный сверхинтеллект или суперинтеллект (Artificial Superintelligence/ASI).

При этом важно отметить, что прикладной ИИ сегодня – это наука, исследования и ИИ-индустрия, ОИИ – это наука и исследования, которые нацелены на получение результата уже в ближайшее время. Что касается суперсильного ИИ, в настоящих условиях – это только гипотетическая концепция. Соответственно тематики прикладного ИИ и ОИИ находятся в центре внимания мирового научного сообщества. Остановимся несколько подробнее на специфике каждого из этих видов.

Прикладной ИИ/AI. Прикладной ИИ предназначен для решения какой-либо одной интеллектуальной задачи, либо их небольшого множества (например, системы для игры в шахматы и Го, распознавания образов, речи и т. д.). Современный ИИ отличается совмещением методов глубинного обучения (Deep Learning), основанных на применении искусственной нейронной сети (ИНС) и технологии больших данных (Big Data). В числе новаций в рамках текущего ИИ аналитики выделяют два новых типа – генеративный ИИ и адаптивный ИИ.

ДИСКУРС. 2025. Т. 11, № 1. С. 16–30
DISCOURSE. 2025, vol. 11, no. 1, pp. 16–30

Генеративный ИИ (Generative AI) представляет собой новую мощную и зрелую разновидность ИИ, которая предназначена для создания новых данных на основе обширного обучающего набора данных, например, изображения, программный код, тексты и др. [6].

В 2022 г. в области генеративного ИИ произошел запуск сразу нескольких ИИ-проектов. В частности, это DALL-E 2, нейронная сеть от компании OpenAI (апрель 2022 г.), Stable Diffusion от Stability AI (август 2022 г.), нейронная сеть Midjourney (от одноименной компании Midjourney, февраль 2022 г.). Эти модели генерируют изображения на основе текстового запроса пользователей и даже коротких фраз, введенных в поиск. Однако текстовый чат-бот ChatGPT-3,5 от компании OpenAI (ноябрь 2022 г.), который называют первой многозадачной и первой универсальной моделью, стал главной новацией 2022 г. В числе наиболее значимых применений ChatGPT – языковой перевод, генерация текста, написание кода, резюмирование текста и др.

Еще более впечатляющим достижением OpenAI в 2023 г. стала версия ChatGPT-4 (март 2023 г.) – первая мультимодальная в линейке GPT. Она способна принимать на вход подсказку и изображения, выдавая текстовый результат, может изображать выбранную роль, говорить в определенном тоне, помогать в решении различных задач и др.

Среди отечественных аналогов можно назвать YandexGPT от компании Яндекс (декабрь 2022 г.), который называют «Chat по-русски». Он умеет решать сложные задачи, отвечает на специализированные запросы от бизнеса и, благодаря пониманию контекста, кратко пересказывает статьи, форматирует текст и пр. [7]. В свою очередь первой российской версией мультимодальной нейросети стал GigaChat от Сбербанка (апрель 2023 г.). Он умеет поддерживать диалог с пользователем, писать программный код, создавать тексты и картинки [8].

Однако компания OpenAI продолжила впечатлять мировую ИИ-общественность своими новациями, в частности:

– ChatGPT-4 Turbo (ноябрь 2023 г.), который может работать с большим количеством информации и «знает» гораздо больше прошлой версии, а также обладает еще несколькими полезными функциями для пользователей и разработчиков. Некоторые из улучшений включают еще и более быстрое время отклика, что делает версию более доступной для различных приложений;

– ChatGPT-4o (май 2024 г.) называют флагманской моделью. Буква «о» в названии расшифровывается как «omni» и указывает на универсальность нейросети. Новая версия значительно расширила свой функционал – чат-бот способен не только воспринимать информацию различных типов одновременно, но и отвечать пользователю, генерируя текст, озвучивая его, создавая изображения и т. д.;

– GPT-4o mini (июль 2024 г.) – уменьшенная версия оригинальной модели. Она сохраняет некоторые ключевые особенности GPT-4o, но предлагает упрощенный функционал.

– GPT-4o1 (сентябрь 2024 г.) – последняя на сегодняшний день версия GPT-4. По оценке OpenAI, o1 обучена «думать», подобно человеку, используя методику обучения с подкреплением. Это значит, что модель анализирует возможные варианты, строит цепочку мыслей (*chain of thought*), проверяет свои шаги и только затем выдает ответ. Такой подход позволяет существенно повысить точность и эффективность модели, что особенно важно в научных приложениях, математике и программировании [9].

ДИСКУРС. 2025. Т. 11, № 1. С. 16–30
DISCOURSE. 2025, vol. 11, no. 1, pp. 16–30

В ближайшее время компания OpenAI обещает представить свою последнюю разработку – GPT-5. Новая языковая модель, по свидетельству генерального директора компании С. Альтмана, может значительно превзойти GPT-4 и его версии [10]. В настоящее время разработка GPT-5 еще не завершена, предполагается, что она станет важной ступенькой на пути к общему ИИ.

Одновременно OpenAI усиливает меры безопасности, совершенствуя внутренние протоколы и взаимодействие с федеральными правительствами, чтобы соответствовать этическим нормам и минимизировать риски.

Таковы некоторые из последних достижений прикладного ИИ, в частности генеративного ИИ/Gen AI, несмотря на ограниченные возможности узких методов.

Общий ИИ. Понятие «сильный искусственный интеллект» было введено американским философом Дж. Сёрлом в 1980 г., впервые охарактеризовавшим AGI так: «…Поскольку подходящим образом программированные компьютеры могут иметь схемы входа и выхода, сходные со схемами входа и выхода у людей, у нас появляется соблазн постулировать у компьютеров ментальные состояния, сходные с человеческими ментальными состояниями» [11], т. е. уровень интеллектуальных возможностей AGI был определен как сравнимый с ЕИ человека.

Термин «общий искусственный интеллект», принятый для описания ИИ, который обладает способностью решать широкий спектр задач на уровне человека и даже превосходить его, появился несколько позднее, а именно в начале 2000-х гг. Автор термина «общий ИИ» – Бен Герцель, известный американский ученый, математик, его называют не только одним из ведущих мировых экспертов в области общего ИИ, но и отцом AGI. «Я придумал этот термин в 2002 или 2003 г., – пишет Б. Герцель, – и каждый год я организовывал конференцию по ОИИ, и за последнее десятилетие мы видели, как концепция растет и процветает довольно плодотворно…» [12].

Сегодня тематика ОИИ стала предметом не только исследовательского поиска современного международного ИИ-сообщества, но и возможностью реализации этого поиска в ближайшем будущем. Так, на открытии саммита «Beneficial AGI Summit 2024» (27.02.24 – 01.03.2024, штат Флорида, США) Бен Герцель поделился с целевой аудиторией информацией о работе своей команды над ранним прототипом системы AGI и представил детальный план развития AGI, ориентированный на то, чтобы технология служила интересам человечества и не оказалась под контролем крупных корпораций или государств. О сроках появления нового ИИ Б. Герцель высказался так: «Я думаю, что к началу 2025 г. у нас может появиться детский AGI. Мы можем назвать его эмбриональным AGI» [13].

В своем выступлении известный ученый также обозначил огромные выгоды AGI, включая освобождение человечества от рутинной работы, окончание всех физических и психических заболеваний, лечение старения и др. В то же время, несмотря на все преимущества, Герцель признает, что AGI может принести с собой ряд проблем, в частности неэтичный AGI, который принесет выгоду только глобальной элите и сделает бедных еще беднее.

Таким образом, сегодня общий ИИ – это уже не только научно-исследовательский поиск международного ИИ-сообщества, но и возможность в ближайшее время обрести практическую реализацию в виде детского AGI, или Baby AGI – раннего прототипа системы

общего ИИ, ориентированного разработчиками в конечном счете на необходимость создания этичного AGI.

Однако если Baby AGI был анонсирован в конце февраля 2024 г., то уже в конце марта 2024 г. три цифровые платформы: SingularityNET/SNET (первая в мире децентрализованная сеть ИИ), Fetch.ai (платформа Web3 для новой экономики ИИ) и Ocean Protocol (децентрализованная платформа обмена данными) объявили о создании альянса Artificial Superintelligence Alliance/Альянс искусственного суперинтеллекта.

Проект возглавили ведущие умы в области децентрализованного ИИ, в частности Б. Герцель – создатель и руководитель SingularityNET/SNET, Хумаюн Шейх – основатель Fetch AI и основной инвестор DeepMind, а также Трент Макконахи – архитектор программного обеспечения на основе ИИ и один из основателей Ocean Protocol [14]. Как высказались сами сооснователи проекта, альянс позволит объединить сильные стороны каждой из платформ, которая сможет бросить вызов монополии технологических гигантов, а также будет развивать децентрализованный сверхинтеллект. При этом подчеркивается, что появление ОИИ, как и в случае с прикладным ИИ, несет с собой не только новые возможности для человека и социума, но и риски, прежде всего этические.

В целом, подводя итоги рассмотрения классификации видов современного ИИ, основанной на сравнении уровня интеллектуальных возможностей человека и ИИ, следует отметить, что каждый из этих видов имеет свои особенности, смысловое значение и тенденции развития. При этом каждый из них отражает специфику как цифровых реалий, так и новой постнеклассической картины мира, которая сформировалась на сегодняшний день и органически включила в себя антропный принцип и информацию в качестве универсальной характеристики мира. В частности, в отличие от неклассической науки «она учитывает соотнесенность знаний об объекте не только с особенностями средств и операций деятельности, но и с ценностно-целевыми структурами» [15].

Актуальность риск-ориентированного подхода. Специалисты отмечают, что сегодня практически все существующие ИИ-системы остаются управляемыми и способны функционировать только с помощью человека. В основе технологии ИНС – самообучение и машинное обучение. Алгоритм способен получать и обрабатывать информацию, после чего на ее основе выстраивать и принимать эффективные и относительно более сложные решения, т. е. современный ИИ может учиться среди прочего и на собственном опыте непосредственно. В связи с тем, что процесс обучения (с подкреплением, с учителем и т. д.) совершенствуется с каждой завершенной задачей, происходит постоянный процесс улучшения и совершенствования ИИ, что, несмотря на ограниченные возможности текущего ИИ, определило возрастающий потенциал его новаций.

Однако именно самообучаемость последних и будущих моделей ИИ увеличивает проблему рисков, при этом эффективного способа, как подчеркивают исследователи, отслеживать контент, созданный ИИ, пока не существует. В связи с этим известный отечественный ученый в области нейронаук, психолиггвистики и теории сознания Т. В. Черниговская, в частности, подчеркивает: «Привычный комментарий, что у ИИ будет только то, чему мы его обучим, – несостоятелен: у эволюции свои законы, и сложные системы любого генеза могут развиваться сами по себе, с малопредсказуемым результатом» [16].

О малопредсказуемых результатах бесконтрольного развития ИИ и необходимости управлять рисками в настоящее время говорят многие как западные, так и отечественные лидеры знаний, отрасли и государств, поэтому сегодня принят к рассмотрению рискоориентированный подход к регулированию ИИ как на международном, так и на национальном уровнях. Однако аналитики отмечают, что период активного бесконтрольного развития ИИ уже подходит к концу, технология сделала первые шаги, когда ей нужна была максимальная свобода, теперь необходимо думать о рисках и безопасности. При этом подчеркивается, что речь идет не о том, чтобы остановить развитие ИИ как высокоинтеллектуальной отрасли, а о необходимости решать проблему управления рисками [5].

Проблемные зоны ИИ этического характера. Проблемы «moral machines» определяются мировым научным сообществом как одни из основополагающих, а их решение относят к области мягкого регулирования. На сегодняшний день в этом направлении уже сделаны важные шаги:

– накоплена определенная правовая и нормативно-этическая база в области этики ИИ, которая постоянно совершенствуется на международном и национальном уровнях [17];

– актуализирована разработка новых, этически обусловленных ИИ-профессий. В частности, озвучиваются такие новые профессии, как менеджер по соблюдению этических норм, специалист по обучению эмпатии, специалист по обучению мировоззрению и локализации, специалист по обучению личностным качествам и др. При этом подчеркивается, что в рамках каждого из этих направлений, потребуются новые функции и новые этически обусловленные навыки [4, с. 151–159];

– отдельная тема в профессиональном IT-сообществе – профессиональная этика разработчиков систем ИИ, а также создание этических кодексов и рекомендаций для ИИ-разработчиков. В зависимости от сферы применения ИИ содержание отдельных кодексов различаются, но в их основе лежит некоторый инвариантный набор ценностных ориентиров, который направлен на безопасность и этичность ИИ по отношению к человеку;

– запущен процесс модернизации образовательной системы, ориентированной на новые подходы к подготовке ИИ-специальностей с позиций человеко-ориентированной цифровой эпохи (Human-centric Digital Age) [18].

Каждое из этих направлений постоянно дорабатывается и совершенствуется. Главное требование – безопасность и этичность современного ИИ.

Так, инструменты мягкого регулирования, а именно документная база, были дополнены первыми в мире правилами для ИИ. Весной 2024 г. Европарламент принял закон «Artificial Intelligence Act», устанавливающий требования и правила для разработчиков ИИ-моделей. Закон вступил в силу 1 августа 2024 г. Правила провозглашают риск-ориентированный подход к ИИ, а также формулируют обязательства для пользователей и разработчиков ИИ, которые учитывают уровень риска используемого ИИ в Европе. В законе выделено четыре категории систем ИИ: с минимальным, ограниченным, высоким и неприемлемым риском [19]. Так, например, закон довольно жестко регулирует генеративный ИИ (в частности, ChatGPT) и «высокорисковые системы, основанные на ИИ», среди которых беспилотные автомобили и медицинское оборудование.

В декабре 2023 г. появилась и такая новация как первый международный стандарт ISO/IEC 42001:2023, который выпустила независимая международная неправительственная

ДИСКУРС. 2025. Т. 11, № 1. С. 16–30
DISCOURSE. 2025, vol. 11, no. 1, pp. 16–30

организация по стандартизации (ISO) [20]. Стандарт является структурированным способом управления рисками и возможностями, связанными с ИИ, т. е. представляет собой, по оценкам аналитиков, определенную матрицу рисков, что позволило разработчикам выработать рекомендации по управлению технологиями ИИ в таких областях как этические соображения, прозрачность и непрерывное обучение.

Таковы в целом основные направления создания инструментов и некоторые из конкретных инструментов «мягкого» регулирования проблемных зон ИИ, находящиеся в центре внимания мирового научного сообщества и ИИ-индустрии. Однако одной из специфических особенностей мягкого регулирования является создание ИИ по образу и подобию ЕИ, хотя на этот счет появились и другие мнения.

ЕИ vs ИИ. Все существующие сегодня версии, представляющие принципы и рекомендации для разработчиков ИИ, ограничивающие их свободу и в написании кода, и в возможности неэтичного использования ИИ-технологий, а также весь этический инструментарий построены на основе аналогии ЕИ и ИИ. В то же время, как подчеркивают эксперты, не все способности ЕИ могут быть формализованы. В частности, профессор В. Финн, известный отечественный специалист в области информатики и управления, подчеркивает: «Идеальный теоретический естественный интеллект – это система знаний, множество интеллектуальных способностей и высшие психические функции, каковыми являются интенция, интуиция, инициатива, воображение и рефлексия» [21]. И если в рамках ЕИ интеллектуальный процесс есть взаимодействие мыслительного и познавательного процессов, то в рамках ИИ – это взаимодействие имитируется и усиливается в интеллектуальных системах (партнерских человеко-машинных системах). Таким образом, на сегодня проблема определения понятия «ИИ» и его видовых проекций связана с вопросами о соотношении ЕИ человека и ИИ, а также оценке их возможностей, включая прояснение границ этих возможностей.

Однако ЕИ человека до сих пор изучен лишь частично и продолжает оставаться для науки одной из величайших загадок Вселенной. Человеческий интеллект – это не только когнитивные и эмоциональные способности, которым сегодня обучают ИИ, но и социальное понимание, принятие этических и правовых решений, культурный код, его смыслы, значения. Эти аспекты интеллекта глубоко укоренены в эволюции, в опыте социального бытия социума и его ценностных оснований. Искусственный интеллект и человеческий мозг представляют собой две разные системы, они не должны конкурировать, а только дополнять друг друга, создавая совершенно новые возможности и перспективы для прогрессивного развития человекоориентированной цифровой эпохи. Отсюда необходимость обращения к анализу ценностных ориентиров современного ИИ, в частности больших языковых моделей LLM (large language model). Их определяют как передовые системы ИИ, которые используют огромные объемы данных и сложные алгоритмы для понимания, интерпретации и создания человеческого языка.

Система ценностей больших языковых моделей/LLM. Последние исследования в области изучения ИИ показали, что при разработке ИИ-систем нельзя проводить прямые аналогии между человеческим и машинным разумом. Называется целый ряд причин, обусловивших вывод: у больших языковых моделей нет индивидуальной базы опыта и знаний, которая используется людьми, LLM недостает эмоциональной гибкости, функциональной смысловой и языковой компетенций и др. [22].

ДИСКУРС. 2025. Т. 11, № 1. С. 16–30
DISCOURSE. 2025, vol. 11, no. 1, pp. 16–30

В частности, исследователи Азиатской научно-исследовательской лаборатории Microsoft и Университета Цинхуа, опираясь на междисциплинарный подход (факторный анализ, лексическая гипотеза, семантическая кластеризация, генеративный подход и др.) смогли выявить как причины, так и набор ключевых ценностей, на которые опираются современные LLM. В частности, выяснилось, что система ценностей современных LLM включает в себя три уровня:

– *компетентность (Competence)*. Утилитарные ценности, связанные с качеством выполнения поставленных задач, для ИИ расположены на первом месте. Это точность, информативность, релевантность и т. д.;

– *характер (Character).* На втором месте расположены социальные и моральные ценности, которые имеют важное значение для человека. Среди них эмпатия, доброта, альтруизм и т. п.;

– *целостность (Integrity).* Такие фундаментальные этические принципы, как справедливость, непредвзятость, конфиденциальность и т. п., расположились ниже в этой иерархии.

В целом это значит, что определяющим критерием оценки ИИ выступают утилитарные ценности, а для человека и социума – нравственные ценности. Отсюда следует вывод, что в коммуникации «человек – ИИ» при принятии решений важно руководствоваться теми рекомендациями, которые учитывают разницу между системой общечеловеческих ценностей и ценностей ИИ, а не только, как это сейчас наиболее распространено, между функциональными возможностями ЕИ и ИИ. В частности, к таким новым рекомендациям можно отнести:

– ИИ только аппроксимирует системы общечеловеческих ценностей, а не использует их;

– системы общечеловеческих ценностей и ценности ИИ совпадают только частично, поэтому проведение аналогии между ЕИ и ИИ, а также их ценностными координатами недопустимо;

– LLM не обладают смысловой языковой компетентностью, которая доступна для человека, несмотря на наличие развитой формальной языковой компетентности (умение правильно применять языковые конструкции и др.);

– для ИИ первостепенными являются связанные с качеством исполнения задач утилитарные ценности: релевантность, точность, информативность, и т. п.;

– и нравственные ценности, и этические принципы имеют абстрактный характер, тем не менее их формализация в оптике взаимодействия «человек – ИИ» занимает важное место в обучении ИИ этическим принципам, на которых строятся партнерские отношения с человеком.

Таким образом, для процесса взаимодействия человека и ИИ необходимо учитывать ценности и принципы, которыми «руководствуется» ИИ для построения эффективной эмоционально-ориентированной коммуникации «человек–ИИ». В этом отношении в качестве более перспективного и западные, и отечественные специалисты рассматривают подход, который получил название «4E Cognition», или «4E».

Подход «4E Cognition». «4E» – это область междисциплинарных исследований, которая строит предположения на том, что в результате взаимодействий между физической и социальной средой, мозгом и телом структурируется и формируется познание. Для этого

подхода «4E» расшифровывается как понимание познания и сознания, как Embodied (телесновоплощенного), Embedded (вписанного в среду), Enacted (связанного с действием) и Extended (расширенного). Подход «4E» постулирует, что познание встраивается, разыгрывается, воплощается или расширяется посредством внечерепных процессов и структур, а не происходит только в голове человека. В частности, известный отечественный ученый, философ, признанный специалист в области философии науки и теории познания В. А. Лекторский отмечает, что подход «исходит из того, что само по себе изучение работы нейронных сетей, сколь бы детальным оно ни было, не может дать ответ на вопрос о природе сознания, ибо последнее определяется не просто работой мозга (хотя без этой работы оно невозможно), а отношением познающего и действующего агента к внешнему миру, включающему как мир природного окружения, так и в случае человека мир, созданный самим человеком – мир культуры» [3, с. 15].

Сегодня подход «4E» рассматривается как наиболее плодотворный во всех отношениях, так как подчеркивает отличие человека от ИИ во всех его существующих и возможных проекциях. При этом, несмотря на то, что ИИ не укладывается под описание 4E-характеристик напрямую, а только опосредованно, этот подход получил распространение в исследованиях, посвященных разработкам ИИ, в первую очередь в робототехнике [23]. Для целого ряда исследователей 4E оценивается как важный методологический инструмент для последующего развития ИИ-отрасли как научного знания и ИИ-индустрии.

В частности, опираясь на концепцию 4E-подхода, специалисты подчеркивают, что ИИ как искусственный объект используется человеком среди прочего для познания мира. В связи с этим он может рассматриваться в качестве Extended AI, т. е. инструмента расширения сознания человека. «Интеллект» ИИ – это в первую очередь результат перенесения на мир вещей, для которого не характерен интеллект, идеи ЕИ, т. е. речь идет о проецировании вовне этого сознания. Также непосредственно воплощенность ИИ является отголоском телесной воплощенности человека – Embodied AI, т. е. датчики роботов служат отсылкой к принципам работы восприятия человека [24].

Исходя из сказанного, особую актуальность приобретает положение канадского философа и культуролога Г. М. Маклюэна, связанное с создаваемыми человеком технологиями и их обратным влиянием на него. Это влияние исследователь рассматривает в контексте мифа о Нарциссе, который оцепенел, когда увидел собственное отражение в воде, т. е. фактически столкнулся с собственным расширением. Маклюэн разбирает такое явление через феномен самоампутации [25, с. 50–52]. При этом делается вывод, что применение и распространение новых технологий выступает источником подобных изменений и способно перестроить мировосприятие человека. «Любое изобретение и любая технология представляют собой внешнюю проекцию, или самоампутацию наших физических тел», – отмечает исследователь [25, с. 54]. Так, человек и его сознание трансформируют среду, что вызывает необходимость подстраиваться под меняющиеся условия и трансформировать восприятие окружающей действительности и собственное восприятие в соответствии с новыми реалиями, а также расширяться при помощи технологий. В результате человек модифицируется и фрагментируется своими технологиями [25, с. 55–56], примером которых выступает современный ИИ.

ДИСКУРС. 2025. Т. 11, № 1. С. 16–30
DISCOURSE. 2025, vol. 11, no. 1, pp. 16–30

В целом аналитики особо подчеркивают продуктивность подхода 4E Cognition по сравнению с подходом, основанном на поиске аналогий между ИИ и ЕИ. Однако представляется, что речь должна идти не просто о сравнении этих подходов, поскольку каждый из них представляет собой разный уровень обобщения и разные целевые установки. В частности, подход *«4E Cognition»* отражает новый уровень понимания единства окружающего мира, в свою очередь, подход, основанный на поиске аналогии между ЕИ и ИИ можно рассматривать как особое направление эволюции науки и исследований ИИ, которое в настоящее время становится моментом доминирующего в ней интеграционного процесса. В новых реалиях процессы дифференциации и интеграции, по оценке специалистов, сливаются в единый интегрально-дифференциальный синтез, который отражает новый уровень понимания единства окружающего мира [26], включая научное знание и исследования в области ИИ. При этом именно благодаря исследовательскому поиску аналогии между ЕИ и ИИ, сегодня подход *«ЕИ vs ИИ»* выступает в качестве одной из важных составляющих подхода 4E, как современного междисциплинарного/конвергентного знания современной когнитивистики. Однако, как подчеркивают специалисты, индустрия и академия (наука/исследования/образование) до сих пор не научились работать в связке.

В частности, молодые ученые из лаборатории Tinkoff Research, по оценкам экспертов, – одна из немногих команд в России, которая занимается научными исследованиями ИИ на базе бизнеса [27], особо отмечают: «Даже крупные игроки почти не работают напрямую с университетами, несмотря на открытые лаборатории и кафедры. Формат, в котором исследователи из академии и инженеры из индустрии находятся в одной рабочей группе, – это большая редкость» [27]. И хотя значимость каждого отдельного направления не подлежит сомнению, сегодня в целях эффективного риск-ориентированного развития этичного ИИ объективируется не только значение совместной деятельности исследователей из академии и инженеров из индустрии, но также необходимость присоединения к этому партнерству и ученых из гуманитаристики в целях создания безопасных и этичных систем как текущего ИИ, так и будущего нового AGI.

Заключение. Определение видов современного ИИ строится на основе разных критериев, а центральное место в классификации видов продолжает занимать подход, основанный на сравнении естественного интеллекта человека и ИИ. Однако, как показали специальные исследования, нельзя проводить аналогии между ЕИ и ИИ и их ценностными координатами: система общечеловеческих ценностей и ценностей ИИ совпадают только частично. При этом важно отметить, что нравственные ценности и этические принципы, которые выстраиваются на этих ценностях, имеют исключительно абстрактный характер. Однако их формализация для взаимодействия «человек – ИИ» имеет важное значение для обучения, а в дальнейшем и самообучения ИИ этическим принципам построения человекоориентированной коммуникации и в целом для дальнейшего прогрессивного развития ИИ-отрасли как научного знания, исследований и ИИ-индустрии. Главный тезис: ИИ не должен быть конкурентом: это сподвижник, ассистент, подручный и одновременно инструмент, который активирован цифровой цивилизацией в помощь и человеку, и социуму в их прогрессивном развитии.

ДИСКУРС. 2025. Т. 11, № 1. С. 16–30
DISCOURSE. 2025, vol. 11, no. 1, pp. 16–30

СПИСОК ЛИТЕРАТУРЫ

1. Зачем искусственному интеллекту этика? / П. М. Готовцев, В. Э. Карпов, Т. А Нестик, Е. Г. Потапова // Этика и «цифра»: этические проблемы цифровых технологий. М.: Изд-во РАНХиГС, 2020. С. 40–42.

2. Культура съедает стратегию на завтрак // TenChat. URL: https://tenchat.ru/media/104720-kultura-syedayet-strategiyu-na-zavtrak (дата обращения: 08.06.2024).

3. Лекторский В. А. Искусственный интеллект в изучении человека, человека в мире, создаваемом искусственным интеллектом // Человек и системы искусственного интеллекта / под ред. В. А. Лекторского. СПб.: Юридический центр, 2022. С. 10–29.

4. Доэрти П., Уилсон Дж. Человек + машина. Новые принципы работы в эпоху искусственного интеллекта / пер. О. Сивченко, Н. Яцюк. М.: Манн, Иванов и Фербер, 2019.

5. Ашманов И. С., Касперская Н. И. Цифровая гигиена. СПб.: Питер, 2021.

6. Banh L., Strobel G. Generative artificial intelligence // Electron Markets. 2023. Vol. 33, № 63. DOI: https://doi.org/10.1007/s12525-023-00680-1.

7. YandexGPT. URL: https://ya.ru/ai/index (дата обращения: 18.09.2024).

8. GigaChat. URL: https://giga.chat/ (дата обращения: 18.09.2024).

9. Марсавин О. OpenAI представила новую модель o1: нейросеть, которая думает и решает в 6 раз лучше GPT-4o // IT World. 12.09.2024. URL: https://www.it-world.ru/tech/9fpmsusqq1c8os8sc00oocwss00occ.html (дата обращения: 28.09.2024).

10. Сэм Альтман: GPT-5 станет в разы более продвинутой версией нынешней GPT-4o // Overclockers. 30.07.2024. URL: https://overclockers.ru/blog/RoadToKnowledge/show/166579/Sem-Al-tman-GPT-5-stanet-v-razy-bolee-prodvinutoj-versiej-nyneshnej-GPT-4o (дата обращения: 28.09.2024).

11. Сёрль Дж. Сознание, мозг и программы. 1980 // Гуманитарный портал. URL: https://gtmarket.ru/library/articles/6661 (дата обращения: 13.09.2024).

12. Gilson C. Ben Goertzel on how blockchain can decentralize artificial intelligence // Cointelegraph. 20.10.2017. URL: https://cointelegraph.com/news/ben-goertzel-on-how-blockchain-can-be-used-to-decentralize-artificial-intelligence (дата обращения: 22.08.2024).

13. Бен Герцель: «Baby AGI» может появиться уже в начале 2025 г. // Securitylab.ru. 29.02.2024. URL: https://www.securitylab.ru/news/546389.php (дата обращения: 13.09.2024).

14. Artificial Superintelligence Alliance. URL: https://www.superintelligence.io (дата обращения: 23.09.2024).

15. Горковенко И. А., Стрельченко В. И. Идея научной картины мира: постнеклассическая рациональность // Вестн. Ленингр. гос. ун-та им. А. С. Пушкина. 2010. Т. 2, № 2. С. 133–141.

16. Черниговская Т. В. Естественный и искусственный интеллект: смыслы или структуры? // Человек и системы искусственного интеллекта / под ред. В. А. Лекторского. СПб.: Юридический центр, 2022. С. 160–171.

17. Мамина Р. И., Ильина А. В. Искусственный интеллект: в поисках формализации этических оснований // Дискурс. 2022. Т. 8, № 6. С. 17–30. DOI: https://doi.org/10.32603/2412-8562-2022-8-6-17-30.

18. Карпова Е. А., Дрынкина Т. И. STREAM-подход как новая образовательная парадигма развития ключевых компетенций XXI века // Вестн. гос. гуманит.-технолог. ун-та. 2022. № 3. С. 35–42.

19. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 // Official J. of the European Union. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689 (дата обращения: 18.09.2024).

20. ISO/IEC 42001:2023. Информационные технологии. Искусственный интеллект. Система менеджмента // Российский ин-т стандартизации. URL: https://www.gostinfo.ru/catalog/Details/?id=7479292 (дата обращения: 18.09.2024).

21. Финн В. Далеко не все функции естественного интеллекта могут быть формализованы и автоматизированы // Коммерсантъ. 23.12.2019. URL: https://www.kommersant.ru/doc/4198609 (дата обращения: 08.08.2024).

22. Кулик А. Раскрывая тайны разума ИИ: уникальные ценности и психологические черты // TenChat. URL: https://tenchat.ru/media/2252628-raskryvaya-tayny-razuma-ii-unikalnyye-tsennosti-i-psikhologicheskiye-cherty (дата обращения: 08.08.2024).

23. Newen A., De Bruin L., Gallagher Sh. The Oxford Handbook of 4E Cognition. Oxford: Oxford Univ. Press, 2018. DOI: https://doi.org/10.1093/oxfordhb/9780198735410.001.0001.

24. Hoffman M., Pfeifer R. Robots as Powerful Allies for the Study of Embodied Cognition from the Bottom Up // The Oxford Handbook of 4E Cognition / A. Newen, L. de Bruinand, Sh. Gallagher. Oxford: Oxford Univ. Press, 2018. P. 841–862. DOI: https://doi.org/10.1093/oxfordhb/9780198735410.013.45.

25. Маклюэн М. Понимание медиа: внешнее расширение человека / пер. с англ. В. Николаева. М.: Жуковский: Канон-пресс-Ц, 2003.

26. Баксанский О. Е. Стратегические цели NBICS-конвергенции: знания, технологии и общество // Россия: тенденции и перспективы развития: сб. статей. 2016. Вып. 11. Ч. II. М.: ИНИОН РАН, 2016. С. 14–19.

27. Наука в ИИ: как российские ученые исследуют искусственный интеллект // ТАСС. URL: https://spec.tass.ru/tinkoff-research/#:~:text=Ученые%20про%20исследования%20искусственного%20инте% (дата обращения: 17.08.2024).

Информация об авторах.

Мамина Раиса Ильинична – доктор философских наук (2007), профессор кафедры философии Санкт-Петербургского государственного электротехнического университета «ЛЭТИ» им. В. И. Ульянова (Ленина), ул. Профессора Попова, д. 5Ф, Санкт-Петербург, 197022, Россия. Автор более 100 научных публикаций. Сфера научных интересов: аксиосфера современного социума, коммуникативные практики, кросскультурное сотрудничество, цифровые коммуникации, цифровой этикет, цифровая самопрезентация, инновационные образовательные траектории.

Ильина Анна Валерьевна – ассистент кафедры философии Санкт-Петербургского государственного электротехнического университета «ЛЭТИ» им. В. И. Ульянова (Ленина), ул. Профессора Попова, д. 5Ф, Санкт-Петербург, 197022, Россия. Автор 10 научных публикаций. Сфера научных интересов: философия культуры, аксиология, этика искусственного интеллекта.

О конфликте интересов, связанном с данной публикацией, не сообщалось.
Поступила 07.10.2024; принята после рецензирования 03.12.2024; опубликована онлайн 20.02.2025.

REFERENCES

1. Gotovtsev, P.M., Karpov, V.E., Nestik, T.A. and Potapova, E.G. (2020), "Why does artificial intelligence need ethics?", *Etika i «tsifra»: eticheskie problemy tsifrovykh tekhnologii* [Ethics and the Digital: Ethical Issues of Digital Technologies], RANEPA; Moscow, RUS, pp. 40–42.

2. "Culture is developing a strategy for breakfast", *TenChat*, available at: https://tenchat.ru/media/104720-kultura-syedayet-strategiyu-na-zavtrak (accessed 08.06.2024).

3. Lektorsky, V.A. (2022), "Artificial Intelligence in Human Studies, the Human being in the World, created by Artificial Intelligence", *Man and systems of artificial intelligence*, Yuridicheskii tsentr, SPb., RUS, pp. 10–29.

4. Daugherty, P. and Wilson, J. (2019), *Human + Machine: Reimagining Work in the Age of AI*, Transl. by Sivchenko, O. and Yatsyuk, N., Mann, Ivanov i Ferber, Moscow, RUS.

5. Ashmanov, I.S. and Kasperskaya, N.I. (2021), *Tsifrovaya gigiena* [Digital Hygiene], Piter, SPb., RUS.

6. Banh, L. and Strobel, G. (2023), "Generative artificial intelligence", *Electron Markets*, vol. 33, no. 63. DOI: https://doi.org/10.1007/s12525-023-00680-1.

7. *YandexGPT*, available at: https://ya.ru/ai/index (accessed 18.09.2024).

8. *GigaChat,* available at: https://giga.chat/ (accessed 18.09.2024).

9. Marsavin, O. (2024), "OpenAI Unveils New Model o1: A Neural Network That Thinks and Decides 6x Better Than GPT-4o", *IT World,* 12.09.2024, available at: https://www.it-world.ru/tech/9fpmsusqq1c8os8sc00oocwss00occ.html (accessed 28.09.2024).

10. "Sam Altman: GPT-5 will be a mich more advanced version of the current GPT-4o" Overclockers. 30.07.2024, available at: https://overclockers.ru/blog/RoadToKnowledge/show/166579/Sem-Al-tman-GPT-5-stanet-v-razy-bolee-prodvinutoj-versiej-nyneshnej-GPT-4o (accessed 28.09.2024).

11. Searle, J. (1980), "Minds, Brains, and Programs", *Gumanitarnyi portal* [Humanitarian portal], available at: https://gtmarket.ru/library/articles/6661 (accessed 13.09.2024).

12. Gilson, C. (2017), "Ben Goertzel on how blockchain can decentralize artificial intelligence", *Cointelegraph*, 20.10.2017, available at: https://cointelegraph.com/news/ben-goertzel-on-how-blockchain-can-be-used-to-decentralize-artificial-intelligence (accessed 22.08.2024).

13. "Ben Goertzel: "Baby AGI could be here as early as 2025" (2024), *Securitylab.ru*, 29.02.2024, available at: https://www.securitylab.ru/news/546389.php (accessed 13.09.2024).

14. *Artificial Superintelligence Alliance*, available at: https://www.superintelligence.io (accessed 23.09.2024).

15. Gorkovenko, I.A. and Strel'chenko, V.I. (2010), "The Idea of a Scientific Picture of the World: "Postnonclassical Rationality", *Pushkin Leningrad State Univ. J.*, vol. 2, no. 2, pp. 133–141.

16. Chernigovskaya, T.V. (2022), "Artificial and Natural Intelligence: Meaning or Structures?", *Man and systems of artificial intelligence*, Yuridicheskii tsentr, SPb., RUS, pp. 160–171.

17. Mamina, R.I. and Ilina, A.V. (2022), "Artificial Intelligence: in Search for Formalization of Ethical Foundations", *DISCOURSE*, vol. 8, no. 6, pp. 17–30. DOI: 10.32603/2412-8562-2022-8-6-17–30.

18. Karpova, E.A. and Drynkina, T.I. (2022), "STREAM technology as a new educational paradigm for the development of key competences of the XXI century", *Vestnik of State Univ. of Humanities and Technology*, no 3, pp. 35–42.

19. "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024", *Official J. of the European Union*, available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689 (accessed 18.09.2024).

20. "ISO/IEC 42001:2023. Information technology. Artificial intelligence. Management system" (2024), *Russian Standardization Institute*, available at: https://www.gostinfo.ru/catalog/Details/?id=7479292 (accessed 18.09.2024).

21. Finn, V. (2019), "Not all functions of natural intelligence can be formalized and automated", *Kommersant*", 23.12.2019, available at: https://www.kommersant.ru/doc/4198609 (accessed 08.08.2024).

22. Kulik, A. "Unlocking the Secrets of the AI Mind: Unique Values and Psychological Traits", *TenChat*, available at: https://tenchat.ru/media/2252628-raskryvaya-tayny-razuma-ii-unikalnyye-tsennosti-i-psikhologicheskiye-cherty (accessed 08.08.2024).

23. Newen, A., De Bruin, L. and Gallagher, Sh. (2018), *The Oxford Handbook of 4E Cognition*, Oxford Univ. Press, Oxford, UK. DOI: https://doi.org/10.1093/oxfordhb/9780198735410.001.0001.

24. Hoffman, M. and Pfeifer, R. (2018), "Robots as Powerful Allies for the Study of Embodied Cognition from the Bottom Up", *The Oxford Handbook of 4E Cognition*, Newen, A., De Bruinand, L. and Gallagher, Sh., Oxford Univ. Press, Oxford, UK. pp. 841–862. DOI: https://doi.org/10.1093/oxfordhb/9780198735410.013.45.

25. McLuhan, M. (2003), *Understanding media: The extensions of man*, Transl. by Nikolaev, V., Kanon-press-C; Kuchkovo pole, Moscow, Zhukovskii, RUS.

26. Backsanskii, O.E. (2016), “Strategic objectives of NBICS-convergence: knowledge, technologies and society”, *Rossiya: tendentsii i perspektivy razvitiya* [Russia: development trends and prospects], iss. 11, part II, INION RAS, Moscow, RUS, pp. 14–19.

27. “Science in AI: How Russian Scientists Research Artificial Intelligence”, *TASS*, available at: https://spec.tass.ru/tinkoff-research/#:~:text=Ученые%20про%20исследования%20искусственного%20инте% (accessed 17.08.2024).

Information about the authors.

Raisa I. Mamina – Dr. Sci. (Philosophy, 2007), Professor at the Department of Philosophy, Saint Petersburg Electrotechnical University, 5F Professor Popov str., St Petersburg 197022, Russia. The author of more than 100 scientific publications. Area of expertise: axiosphere of modern society, communication practices, cross-cultural cooperation, digital communications, digital etiquette, digital self-presentation, innovative educational trajectories.

Anna V. Ilina – Assistant Lecturer at the Department of Philosophy, Saint Petersburg Electrotechnical University, 5F Professor Popov str., St Petersburg 197022, Russia. The author of 10 scientific publications. Area of expertise: philosophy of culture, axiology, ethics of artificial intelligence.

No conflicts of interest related to this publication were reported.
Received 07.10.2024; adopted after review 03.12.2024; published online 20.02.2025.