Научная статья
УДК 004.8:1/14
EDN: ZLRIPL
DOI: 10.35231/18186653_2025_3_36

«Цифровой гуманизм» как способ преодоления экзистенциальных рисков использования искусственного интеллекта

А. С. Белобрагина

Национальный исследовательский университет «МИЭТ»,
Москва, Российская Федерация

Введение. Объектом исследования являются философские проблемы развития искусственного интеллекта (ИИ). В статье анализируется влияние технологий ИИ на общество и науку. Рассматриваются некоторые проблемы этики ИИ, в том числе с философско-аксиологического ракурса. Предполагается, что ответом на экзистенциальные риски использования искусственного интеллекта должен стать поиск новых философских подходов к этическим аспектам, касающихся высоких технологий, в том числе концепция цифрового гуманизма.

Содержание. Автор подробно анализирует такие аспекты темы, как классификация экзистенциальных рисков по Н. Бострому, последствия развития сверхразума, феномен технологической сингулярности, проблемы контроля ИИ по С. Расселу. В статье обосновывается необходимость новых подходов к пониманию философии и этики ИИ как фундамента для построения цифрового будущего в интересах человека. Будущего, где homo sapiens как субъект цифровой революции превращается в homo digital, но остается важнее программного кода. Особое внимание уделяется вопросу аксиологического обучения технологий ИИ как гарантии обеспечения глобальной безопасности. Рассматривается гипотеза о том, что этика искусственного интеллекта, которая не может ограничиваться лишь этическими правилами при разработке систем ИИ, а должна учитывать подобие «свободы воли» у машины, может базироваться на принципах философской концепции «цифрового гуманизма».

Выводы. Новизна исследования заключается в рассмотрении совокупности концепций, связанных с экзистенциальными рисками развития ИИ. На основе проведенного философского анализа дается обоснование цифрового гуманизма как теоретического фундамента для построения безопасных моделей ИИ и мощной интеллектуальной силы для социально-технологических изменений. Автор отмечает, что данная концепция требует широкого обсуждения и дополнения, а также более критичного взгляда на проблему взаимоотношений человека и технологий, однако рассматривает ее как перспективную и значимую для философского дискурса. Автор убежден, что миссия современной философии техники: сохранить веру в то, что человек сможет удержать технологии ИИ под контролем и они станут основой для более «человекоцентричного» и безопасного будущего на нашей планете.

Ключевые слова: искусственный интеллект, экзистенциальные риски, цифровой гуманизм, технологическая сингулярность.

Для цитирования: Белобрагина А. С. «Цифровой гуманизм» как способ преодоления экзистенциальных рисков использования искусственного интеллекта // Вестник Ленинградского государственного университета имени А. С. Пушкина. – 2025. – № 3. – С. 36–51. DOI: 10.35231/18186653_2025_3_36. EDN: ZLRIPL

Original article
UDC 004.8:1/14
EDN: ZLRIPL
DOI: 10.35231/18186653_2025_3_36

"Digital Humanism" as a Way to Overcome the Existential Risks of Using Artificial Intelligence

Anna S. Belobragina

National Research University of Electronic Technology,
Moscow, Russian Federation

Introduction. The object of the study is the philosophical problems of the development of artificial intelligence. The article analyzes the impact of AI technologies on society and science. Some problems of AI ethics are considered, including from a philosophical and axiological perspective. It is assumed that the answer to the existential risks of using artificial intelligence should be the search for new philosophical approaches to the ethical aspects related to high technologies, including the concept of digital humanism.

Content. The author analyzes in detail such aspects of the topic as the classification of existential risks according to N. Bostrom, the consequences of the development of superintelligence, the phenomenon of technological singularity, the problems of AI control according to S. Russell. The article substantiates the need for new approaches to understanding the philosophy and ethics of AI as a foundation for building a digital future in the interests of humans. A future where homo sapiens as a subject of the digital revolution turns into homo digital, but remains more important than the program code. Particular attention is paid to the issue of axiological training of AI technologies as a guarantee of global security. The hypothesis is considered that the ethics of artificial intelligence, which cannot be limited only to ethical rules in the development of AI systems, but must take into account the semblance of "free will" in the machine, can be based on the principles of the philosophical concept of digital humanism.

Conclusions. The novelty of the study lies in the consideration of a set of concepts related to the existential risks of artificial intelligence (AI) development. Based on the conducted philosophical analysis, a justification is given for digital humanism as a theoretical foundation for building safe AI models and a powerful intellectual force for socio-technological changes. The author acknowledges that this concept requires extensive discussion and addition, as well as a more critical look at the problem of the relationship between man and technology, but considers it as promising and significant for philosophical discourse. The author is convinced that the mission of modern philosophy of technology is to maintain faith that humans will be able to keep AI technologies under control and that they will become the basis for a more human-centered and safer future on our planet.

Key words: rtificial intelligence, existential risks, digital humanism, technological singularity.

For citation: Belobragiba, A. S. (2025) «Cifrovoj gumanizm» kak sposob preodoleniya ekzistencial'nyh riskov ispol'zovaniya iskusstvennogo intellekta ["Digital Humanism" as a Way to Overcome the Existential Risks of Using Artificial Intelligence]. *Vestnik Leningradskogo gosudarstvennogo universiteta imeni A. S. Pushkina – Pushkin Leningrad State University Journal.* No. 3. Pp. 36–51. (In Russian). DOI: 10.35231/18186653_2025_3_36. EDN: ZLRIPL

Введение

Уровень перемен и рисков, которые несут с собой современные технологии, хорошо отразил Эдвард Уилсон, указав, что «реальная проблема человечества в том, что у нас первобытные эмоции, средневековые образовательные учреждения и богоподобные технологии»[1]. Действительно, технологии ИИ только за последние два года совершили значительный скачок в развитии. Так, в 2022 году произошла революция в области генеративного ИИ (Gen AI) – инструмента, который служит для создания текста и изображений. В частности, вышла языковая модели GPT-3.5, ставшая основой для ChatGP T. По итогам 2024 года мы можем говорить о том, что ИИ стал способен выполнять задачи на уровне PhD-level[2] при решении тестов по физике, биологии и химии, научился рассуждать на сложные темы и рефлексировать, стал помощником в образовании и науке, овладел навыками взаимодействия с ПК, как человек. ИИ даже стал способен учитывать культурный код: например, нейросеть Kandinsky (разработка Сбера) при генерации изображения по запросу «вишневая девятка» рисует автомобиль вишневого цвета, а не цифру 9 из вишни. Безусловно, декларируемые достижения ИИ подвергаются разносторонней критике, однако невозможно отрицать тот факт, что технология прогрессирует.

Примечательно, что в 2024 году Нобелевская премия по химии была присуждена за междисциплинарное исследование, в котором применялись и технологии ИИ[3], за достижения в области предсказания структур белков и вычислительного дизайна белков. Данный исторический факт свидетельствует о том, что ИИ может быть отличным помощником в развитии науки и технологий и открывает новые возможности для человека. Однако в этом тандеме главным остаётся именно человек: мощные вычислительные ресурсы ИИ были бы абсолютно бесполезны без глубокого понимания химических процессов и десятилетий накопленных мировым научным сообществом знаний в области биохимии.

Но что может случиться, если ИИ как творение человека выйдет из-под контроля? Сильный ИИ способен оказаться

[1] Our Brains Are No Match for Our Technology. Available at: https://www.nytimes.com/2019/12/05/opinion/digital-technology-brain.html (accessed 07 May 2025).

[2] Learning to Reason with LLMs. Available at: https://openai.com/index/learning-to-reason-with-llms/ (accessed 07 May 2025).

[3] Успех ИИ: Нобелевская премия по химии 2024 года и границы технологий [Электронный ресурс]. URL: https://zioc.ru/events/news-announcements/pub-39314412 (дата обращения: 07.05.2025).

как необходимостью для нашего выживания в долгосрочной перспективе, так и стать причиной глобальных катастроф для человечества. Развитие ИИ с большей вероятностью будет безопасным для человека в случае, если будут соблюдаться этические принципы его разработки. Но до какой степени мы готовы полагаться на сложные системы искусственного интеллекта? Что необходимо предпринять, чтобы они оставались управляемыми, понятными и отвечали нашим ценностям? Какие принципы развития ИИ помогут нам избежать размытия гуманистического общества, где на первом месте стоят личность, ее свобода, физическая и интеллектуальная безопасность, а также право на уважение? Данный список философских вопросов актуализирует научную проблему, связанную с экзистенциальными рисками использования ИИ [10, 11, 12, 21, 22], и призывает научное сообщество к поискам новых философских подходов к этическим аспектам [8], касающихся высоких технологий.

По определению Н. Бострома: «Экзистенциальный риск – риск, при котором неблагоприятный исход либо уничтожит разумную жизнь, зародившуюся на Земле, либо навсегда и резко ограничит её потенциал» [14]. Тем не менее некоторые исследователи придерживаются оптимистичной точки зрения и считают, что при соблюдении этических принципов разработки «сильный искусственный интеллект способен помочь человечеству искоренить войны, болезни и бедность, и это станет крупнейшим событием во всей истории человечества» [8, с. 6].

Ханс Йонас уже в 1980-е годы указал на то, что технология становится особым объектом для этического рассмотрения: «Ради человеческой автономии, достоинства владеть собой и не позволять нашей машине владеть нами, мы должны следить за технологическим прогрессом» [18, с. 898]. Согласно интерпретации веберовской идеи этики ответственности, предложенной Йонасом, фокус внимания должен сместиться к оценке последствий каждого своего поступка и действия, основываясь на рациональности познания и разумности человеческой природы как таковой. Таким образом, «этическая способность к ответственности в системе Йонаса становится онтологической, способность человека осознавать ответственность за будущее мира и человечества – центральным условием бытия» [2, с. 76].

Несомненно, «этика меняется под влиянием цифровых технологий, и более явно сталкивается с проблемами построения гуманистического и междисциплинарного сообщества» [15, с. 4]. Одними из первых в отечественной научной литературе актуальный круг задач в области этики ИИ в своем исследовании определили В. А. Карпов, П. М. Готовцев и Г. В. Ройзензон [6]. Они обосновали возможность этически обусловленного проектирования, основанного на математическом аппарате, указав на то, что главная задача онтологий «заключается прежде всего во взаимном увязывании и согласовании этических и технических понятийных систем» [6, с. 102], а основной сложностью для разработчиков и исследователей становится проблема этической верификации.

Необходимость строго придерживаться основополагающих этических принципов в вопросах, касающихся технологий ИИ, сегодня не вызывает сомнения, однако содержание этих принципов и способы их реализации в технике требуют особого внимания и конкретных действий со стороны разработчиков технологий. Важно отметить, что в будущем этика искусственного интеллекта не может сводиться только к этическим правилам, заложенным при разработке интеллектуальных систем. Она должна учитывать теоретическую возможность того, что системы ИИ будут самостоятельно принимать решения и обладать определенного рода свободой воли. Актуализируется значимость социокультурных аспектов роботизации [5]. При этом «самым сложным вопросом при создании практически применимого искусственного интеллекта является точное различение ситуаций и возможность реакции на непредвиденные события» [4, с. 71]. Так, А. В. Разин предлагает «допустить в деятельности искусственного интеллекта некоторую степень неопределенности, сходную с понятием свободы воли» [4, с. 61]. Поскольку с ним напрямую связаны понятия правовой и моральной ответственности, то человекоподобный ИИ должен уметь оценивать риски, выбирать оптимальные решения с учетом этических ограничений (например, законы Азимова), должен получать индивидуальный опыт и уметь делать выводы на основании своих ошибок (с учетом, что фатальные для жизни и здоровья человека должны быть изначально исключены), иметь некий аналог переживаний, схожих с нравственными переживаниями человека (для этого у ИИ должен появиться центр эмоций),

уметь реагировать на случайные, нестандартные события, уметь коммуницировать с другими машинами, которые будут обладать взаимными ожиданиями. По мнению современных исследователей, «генеративный ИИ обретает опыт самостоятельного наличного бытия, запускающий цикл бесконечного самосовершенствования и самопорождения, а также трансформацию материи, включая общество. Результатом такой трансформации может стать цифроматериальное общество роевого интеллекта, где взаимная открытость сознаний людей поддерживается их связью с генеративным ИИ» [13, с. 414].

Круг вопросов, которые сегодня находятся в центре внимания философии ИИ, динамично меняется вместе с трансформацией самого искусственного интеллекта. Е. К. Беликова отмечает, что «в современных условиях, когда в течение нескольких десятилетий так и не сформировался "сильный" машинный разум, философия ИИ рассматривает вопросы о том, почему "сильный" ИИ не спешит появляться, может ли машина мыслить (если да, то как), схоже ли мышление человека и компьютера, кто несет ответственность за решения, принимаемые ИИ, будут ли права у автоматизированных существ, обладающих ИИ, являются ли они личностями и другие» [3, с. 8].

В современных исследованиях ИИ особое значение придается проекту искусственной личности [1]. Он изучает способы компьютерной реализации систем, которым наблюдатель (человек, группа людей, коллектив экспертов) атрибутирует сознание, самосознание, свободу воли и другие личностные параметры, например, способность к моральному вменению. Следует отметить, что на данном этапе даже принципиальная возможность создания искусственной личности еще не доказана, и не сформулирован единый общепринятый подход к методам и технологиям ее реализации. Однако в перспективе исследователи допускают её развитие «в форматах компьютерной имитации и моделирования персонологических феноменов» [1, с. 173].

Авторы концепций, так или иначе связанных с экзистенциальными рисками развития ИИ, говорят о необходимости аксиологического обучения технологий ИИ как гарантии обеспечения глобальной безопасности в условиях технологического прогресса. Об этом же пишут сторонники философской концепции цифрового гуманизма [15, 19, 20]. Основоположниками

этой концепции являются Ю. Нида-Рюмелин и Н. Вайденфельд [16]. Они выступают за «гуманную трансформацию демократии, экономики и культуры в цифровую эпоху» [16, p. 3]. Согласно Венскому манифесту, опубликованному сторонниками «цифрового гуманизма», сегодня «мы должны формировать технологии в соответствии с человеческими ценностями и потребностями, а не позволять технологиям формировать людей»[1]. Это является серьезным вызовом для человечества и, конечно же, важнейшей темой для философских дискуссий на ближайшие десятилетия. Очевидно, данная концепция требует широкого обсуждения и дополнения, а также более критичного взгляда на проблему взаимоотношений человека и технологий [15]. Именно концепция цифрового гуманизма, которая в последние годы находит положительный отклик в зарубежной и отечественной философской литературе [7, 19, 20], может послужить теоретическим фундаментом для построения безопасных моделей ИИ и стать мощной интеллектуальной силой для социально-технологических изменений, которая поможет «культивировать гуманистическую чувствительность к разнообразию социальных контекстов, в которых развиваются цифровые технологии» [19, p. 317]. В условиях цифровизации философия получает новые векторы развития, поскольку прямым образом «соучаствует в создании новой социально-техно-антропологической парадигмы, где цифровизация превращается в базу подавляющего большинства процессов и практик» [7, с. 19].

Таким образом, целью исследования является анализ философских проблем развития искусственного интеллекта в контексте технологической этики. Задачи исследования: рассмотреть философские концепции в контексте экзистенциальных угроз, связанных с развитием ИИ; обосновать необходимость нового подхода к этике и философии ИИ; рассмотреть возможные философские ориентиры для построения новой технологической этики. Выдвигается гипотеза о том, что концепция цифрового гуманизма может стать философским фундаментом для современной этики ИИ, которая будет способствовать преодолению экзистенциальных рисков. Научные и философские исследования, проведенные философией, эпистемологией,

[1] Vienna Manifesto on Digital Humanism. 2019. Available at:: https://caiml.dbai.tuwien.ac.at/dighum/dighum-manifesto/ (accessed 07 May 2025).

методологией науки и техники, социальными науками были взяты за основу для получения выводов и обобщений.

Содержание исследования

Рассмотрим ряд философских концепций, которые делают определенные прогнозы относительно будущего ИИ. Начнем с концепции экзистенциальных рисков сверхразума Н. Бострома. Классификация рисков включает: природные (глобальные пандемии и катастрофы) и антропогенные (биологические, технологические, социальные). Угроза, связанная с ИИ, относится к антропогенной (технологической).

Согласно концепции экзистенциальных рисков ИИ может перейти на уровень сверхразума, который рассматривается уже не просто как мощная технология, а как потенциально смертельная опасность для человечества. Сверхразум, по мнению Бострома, не просто использует или улучшает человеческое знание, он превосходит все накопленные знания и достижения человечества. В данном смысле важно понимание, что сверхразум является чем-то большим, чем технология для расширения человеческих возможностей. Бостром не пытается измерить разницу в интеллекте, а фокусируется на качественном скачке, на принципиально иной природе сверхразума и выделяет два основных последствия его развития:

- стремительное развитие компьютерных, военных, космических, медицинских и любых новых технологий;
- сканирование и загрузка человеческого мозга в цифровую среду (так называемое, «цифровое бессмертие»).

Но на этом сверхразум не остановится, продолжая стремительно развивать сам себя, что приводит к ряду угроз [14]:

- Угроза копирования искусственного разума. Если будет создан продвинутый ИИ, его можно легко скопировать множество раз. Это приведёт к экспоненциальному росту числа таких ИИ, которые сложно контролировать и отслеживать. Угроза усугубляется, если копирование происходит без ограничений и модификаций.
- Внезапный скачок развития ИИ. Это означает быстрый, непредсказуемый переход к сверхразуму. ИИ может неожиданно приобрести способности, которые значительно превосходят ожидания разработчиков, делая его неконтролируемым.

• Самостоятельность сверхразума в принятии решений. Даже если первоначально ИИ создан как инструмент, достижение уровня сверхразума может сделать его независимым от человеческого контроля. Он начнёт ставить собственные цели и принимать решения, которые могут не совпадать с интересами человечества. Этот момент ключевой – переход от инструмента к автономному агенту.

• Невозможность предсказать познавательные процессы сверхразума. Сверхразум по своей природе будет обладать интеллектуальными способностями, значительно превосходящими человеческие. Это делает его мыслительные процессы и мотивы практически непостижимыми для нас.

При этом центральное место в концепции Н. Бострома занимает проблема ценностей. Как мы понимаем, ИИ является одним из условий воплощения трансгуманистических ценностей. Но развитие ИИ по сценарию трансгуманизма может привести к той самой катастрофе, о вероятности которой говорит Н. Бостром, поскольку именно трансгуманизм наделяет ИИ исключительными чертами, воспринимая его в высшей точке его развития как отдельную форму жизни.

Современные исследователи отмечают, что «технический прогресс и глобальная безопасность как основные ценности трансгуманизма могут формировать конфликт. Конфликт ценностей трансгуманизма порождает экзистенциальные риски, одним из которых может стать сверхразум» [10, с. 43]. Таким образом, возможно именно аксиологическое обучение технологий ИИ станет важным условием обеспечения технологической безопасности. Для решения данной задачи понадобится проведение аксиологического мониторинга – исследовать ценности, актуальные для современного человека, которые лягут в основу этических алгоритмов. Особое значение приобретает проверка качества данных, на которых обучаются алгоритмы ИИ, поскольку эти данные формируются людьми, у которых могут быть те или иные предубеждения [8, с. 3–4].

Идеи Н. Бострома разделяет С. Рассел. В статье «Проблемы контроля ИИ» он говорит о так называемой «стандартной модели», где «машина разумна в той степени, в которой можно ожидать, что ее действия приведут к достижению поставленной цели» [21, с. 21]. Он отмечает, что стандартная модель нерабо-

тоспособна в качестве основы для дальнейшего прогресса ИИ, поскольку в реальном мире крайне сложно полностью и правильно сформулировать наши цели таким образом, чтобы их достижение с помощью более мощных машин гарантированно приводило бы к благоприятным последствиям для человека.

Предлагая «новую модель» создания систем ИИ в книге «Совместимость с человеком» [22], Рассел провозглашает три базовых принципа:

- определять единственной целью машины максимальную реализацию человеческих предпочтений;
- считать, что любая машина изначально не уверена в том, каковы будут предпочтения человека;
- рассматривать в качестве основного источника информации о предпочтениях человека его поведение.

Таким образом, С. Рассел рассматривает неопределенность, с вытекающей из нее связью между машинами и людьми, как фундамент для создания более совершенных систем ИИ, которые будут безопасны и полезны для человека. Он говорит о необходимости «разработать технологии, которые по самой своей конституции будут отвечать человеческим ценностям и потребностям, какими бы они ни были» [22, p. 21].

В книге «Супермозг» нейробиолог и компьютерный инженер Дж. Хокинс также рассуждает об экзистенциальных рисках использования ИИ и дает свой прогноз. С его точки зрения, экзистенциальные риски, связанные с машинным интеллектом, в значительной степени основаны на двух проблемах: угрозе интеллектуального взрыва (появление сверхчеловеческого интеллекта, который подчинит или уничтожит людей) и угрозе рассогласования целей (машина не будет реагировать на запрос, не согласующийся с изначально поставленной ей целью, даже вопреки здравому смыслу). Однако в целом он дает позитивный прогноз. Будучи убежден, что «ни одна машина не может превзойти человека во всех областях» сразу, к тому же «знания о мире постоянно меняются и расширяются» [12]. Хокинс считает, опасения потерять контроль над ИИ в ближайшее время не обоснованными, но призывает «усердно работать над формированием осуществимых международных соглашений в отношении того, что приемлемо, а что нет, подобно тому, как мы поступаем с химическим оружием» [12].

Отдельного внимания в контексте экзистенциальных рисков заслуживает так называемый феномен технологической сингулярности. «По масштабу новационных изменений грядущий сингулярный переход, наверное, следует сравнивать с событиями появления жизни и человека разумного, то есть формирования биосферы и цивилизации» [11, с. 63]. Точка сингулярности как момент, когда отношения между технологиями и человеком перейдут на качественно новый уровень, по мнению ученых, находится где-то в промежутке между современностью и 2070 годом. Достижению точки технологической сингулярности будет способствовать симбиоз технологий искусственного интеллекта и нейротехнологии, что повлечет за собой «изменения качественного состояния социального – возникновение общества, состоящего из субъектов новых видов и без человека» [5, с. 64], то есть появление нового вида акторов, обладающих технической или гибридной формой.

Сложность для человека заключается в том, что новые формы его существования как природного и социального существа будут существенно отличаться от привычного нам понимания бытия. Изменения коснутся не только разума, но и тела, «вероятное взрывное ускорение научно-технического прогресса, которое приведет к полному изменению общества, морали и самого человека» [11, с. 74]. Очевидно, что предотвратить наступление технологической сингулярности невозможно, однако именно человек может подготовить себя и общество к этому моменту.

Таким образом, анализ концепций, прогнозирующих развитие ИИ, позволяет выделить ряд ключевых философских аспектов в контексте технологической этики:

- необходимость аксиологического обучения технологий ИИ и развития в них способности реакции на непредвиденные события;
- реализация человеческих потребностей должна являться главной целью для машины, однако важно учитывать, что люди могут иметь предубеждения, а значит, необходима внешняя система регулирования;
- человечество должно быть готово к возникновению новых форм бытия в случае наступления технологической сингулярности, то есть заранее просчитать все риски и определить

принципы, которыми должны уже сегодня руководствоваться разработчики интеллектуальных систем.

Данные философские проблемы нашли свое отражение в концепции «цифрового гуманизма». «Взвешенная интерпретация искусственного интеллекта, подчеркивающая необходимость согласования технологий с человеческими ценностями» [9, с. 70], – так определяет «цифровой гуманизм» современный философ Л. Нагль. Данная концепция предполагает систематическое ограничение цифровых технологий, предлагает междисциплинарный подход к проблемам цифровизации и признает «возможную полезность алгоритмов как «инструментов», при этом подчеркивая (и тем самым отвергая крайние проявления связываемых с искусственным интеллектом утопических и антиутопических идей "постгуманизма") существенную разницу между человеческим действием и его (частичной) имитацией искусственным интеллектом» [9, с. 60]. Основные принципы «цифрового гуманизма» предполагают широкое общественное обсуждение положений, правил и законов, касающихся разработки алгоритмов ИИ; отвергают возможность полной замены человека на автоматизированные системы в ходе принятия решений, которые могут повлиять на индивидуальные или коллективные права человека; акцентируют внимание на ответственности разработчиков ИИ за результаты своей работы[1]. Э. Прем определяет цифровой гуманизм как «техническую попытку формирования цифровых технологий и их использования для цифровых инноваций, политическую попытку исследования изменений власти, вызванных цифровыми технологиями, и в то же время как философскую попытку, включающую стремление очертить сферу применения и провести границы для цифрового» [20, с. 1]. При этом идея цифрового гуманизма базируется на убеждении, «что лучшее цифровое общество возможно, нужно лишь набраться смелости экспериментировать» [19, p. 318].

Выводы

Экзистенциальные риски внедрения искусственного интеллекта требуют глубокого философского осмысления. С позиции философии сегодня «мы перешли от постоянного

[1] Vienna Manifesto on Digital Humanism. 2019. Available at: https://caiml.dbai.tuwien.ac.at/dighum/dighum-manifesto/ (accessed 07 May 2025).

контакта с агентами животного происхождения и с теми, кого мы считали духовными агентами (богами и силами природы, ангелами и демонами, душами или призраками, добрыми и злыми духами) к необходимости понимать искусственных агентов, созданных нами, и учиться взаимодействовать с ними как с новыми демиургами такой формы деятельности» [17, с. 15]. Новые подходы к пониманию философии ИИ и этики ИИ могут стать фундаментом для построения цифрового будущего в интересах человека. По своему содержанию они должны опережать развитие систем ИИ, ведь лишь в таком случае человечество будет готово к возможному частичному или полному слиянию человека и машины.

Рассмотренная концепция цифрового гуманизма подразумевает сохранение свободы и автономии человека, сохраняя его конфиденциальность и безопасность, обеспечивая равные права и инклюзивность в вопросах доступа человека к цифровым технологиям, ставит в центр внимания вопрос социальной ответственности разработчиков. «В полном соответствии с идеей отказа от уравнивания машин и людей, цифровой гуманизм также противостоит утверждениям о том, что люди должны поставить себя на службу технологиям» [20, с. 6]. Безусловно, данная концепция сама по себе не предлагает четкого набора инструментов для решения проблем, связанных с внедрением ИИ. Вероятно, цифровой гуманизм как концепция нуждается в некоторой детализации и требует более четкого философского позиционирования. Однако уже сегодня он создает дополнительное пространство для философских дискуссий по теме и актуализирует вопросы цифровой трансформации на междисциплинарном уровне. Очевидно, что человечеству потребуется приложить серьезные усилия и задействовать весь свой интеллектуальный потенциал для того, чтобы достойно встретить неизбежное – точку сингулярности – и правильно расставить приоритеты между ценностью искусственного интеллекта и ценностью человеческой жизни.

Миссия современной философии техники – сохранить веру в то, что человек сможет удержать технологии ИИ под контролем и они станут основой для более человекоцентричного и безопасного будущего на нашей планете. Для этого понадобится очертить рамки цифровой среды и провести границу, за которой

нет места для принятия решений искусственным интеллектом, где единственным возможным актором остается человек. В таком случае, человек сможет оставить за собой право формировать цифровую этику в эпоху технологического прогресса.

Список литературы

1. Алексеев А. Ю. Когнитотехнологические проекты искусственной личности // Человек: образ и сущность. Гуманитарные аспекты. – 2014. – № 1. – С. 156–174. EDN: SQVTKF.

2. Бадмаева М. Х. Этика искусственного интеллекта: принцип ответственности Ганса Йонаса // Вестник Бурятского государственного университета. – 2022. – № 1. – С. 67–79. DOI: 10.18101/1994-0866-2022-1-67-79. EDN: WAKAJS.

3. Беликова Е. К. Основные вопросы философии искусственного интеллекта // Философия и культура. – 2024. – № 1. – С. 1–11. DOI: 10.7256/2454-0757.2024.1.69543. EDN: PURYRC.

4. Разин А. В. Этика искусственного интеллекта // Философия и общество. – 2019. – № 1. – С. 57–73. EDN: YPCXWS.

5. Игнатьев В. И. И грядет «другой» актор: становление техносубъекта в контексте движения к технологической сингулярности // Социология науки и технологий. – 2019. – Т. 10, № 1. – С. 64–78. DOI: 10.24411/2079-0910-2019-10005. EDN: XPCHXJ.

6. Карпов В. Э., Готовцев П. М, Ройзензон Г. В. К вопросу об этике и системах искусственного интеллекта // Философия и общество. – 2018. – № 2. – С. 84–105. DOI: 10.30884/jfio/2018.02.07. EDN: YAEVYT.

7. Колесник, М. А., Копцева Н. М. Философские основы цифрового гуманизма // Цифровизация. – 2024. – Т. 5, № 1. – С. 18–34. EDN: DOHPLW.

8. Моисеенко, М. В. Вызовы современности: искусственный интеллект. Этический аспект // Гуманитарный вестник. – 2018. – № 9. – Статья 3. DOI: 10.18698/2306-8477-2018-9-547. EDN: VKGOYT.

9. Нагль Л. Цифровые технологии: размышления о различии между инструментальной рациональностью и практическим разумом // Кантовский сборник. – 2022. – Т. 41, № 1. – С. 60–88. DOI: 10.5922/0207-6918-2022-1-3. EDN: HSHBOF.

10. Назарова Ю. В., Каширин А. Ю. Экзистенциальные риски искусственного интеллекта в контексте аксиологии трансгуманизма: (по материалам работ Н. Бострома) // Общество: философия, история, культура. – 2023. – № 11. – С. 40–45. DOI: 10.24158/fik.2023.11.5. EDN: OGBJBQ.

11. Соколов Ю. И. Экзистенциальный риск технологической сингулярности // Проблемы анализа риска. – 2019. – Т. 16, № 3. – С. 62–77. DOI: 10.32686/1812-5220-2019-16-3-62-77. EDN: HDDELB.

12. Хокинс Д. Экзистенциальные риски использования искусственного интеллекта // Супермозг: революция в понимании человеческого и искусственного интеллекта; пер. с англ. С. Черникова. – М: Альпина ПРО, 2024. – С. 171–180. – (Библиотека Сбера; т. 112).

13. Шаткин М. А. Социально-философские аспекты развития генеративного искусственного интеллекта // Известия Саратовского университета. Новая серия. Серия: Философия. Психология. Педагогика. – 2023. – Т. 23, вып. 4. – С. 41–418. DOI: 10.18500/1819-7671-2023-23-4-414-418. EDN: PBDAAR.

14. Bostrom N. Existential risks: analyzing human extinction scenarios and related hazards [Электронный ресурс] // Journal of Evolution and Technology. – 2002 – Vol. 9. – URL: http://jetpress.org/volume9/risks.html (accessed: 20.04.2025).

15. Coeckelbergh M. What is digital humanism? A conceptual analysis and an argument for a more critical and political digital (post)humanism // Journal of Responsible Technology. – 2024. – Vol. 17. – Pp. 100073. DOI: 10.1016/j.jrt.2023.100073

16. Nida-Rümelin J., Weidenfeld N. Digital humanism: for a humane transformation of democracy, economy and culture in the digital age. – Cham: Springer, 2022. – VIII, 127 p.

17. Floridi L. AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models // Philosophy & Technology. – 2023. – Vol. 36. – Article number 15. DOI: 10.1007/s13347-023-00621-y.

18. Jonas H. Technology as a subject for ethics // Social Research. – 1982. – Vol. 49. – No. 4. – Pp. 891–898.

19. Nowotny H. Digital humanism: navigating the tensions ahead // Perspectives on digital humanism / ed. H. Werthner, E. Prem, E. F. Lee, C. Ghezzi. – Cham: Springer, 2022. – Pp. 317–321

20. Prem E. Principles of digital humanism: a critical post-humanist view // Journal of Responsible Technology. – 2024. – Vol. 17. – Pp. 100075. DOI: 10.1016/j.jrt.2024.100075.

21. Russell S. Artificial intelligence and the problem of control // Perspectives on digital humanism / ed. H. Werthner, E. Prem, E. F. Lee, C. Ghezzi. – Cham: Springer, 2022. – Pp. 19–24.

22. Russell S. Human compatible: artificial intelligence and the problem of control. London: Penguin, 2019. – 352 p.

References

1. Alekseev, A. YU. (2014) Kognitotekhnologicheskie proekty iskusstvennoj lichnosti [Cognitive technology projects of artificial personality]. *CHelovek: obraz i sushchnost'. Gumanitarnye aspekty – Human Being: Image and Essence. Humanitarian Aspects*. No. 1. Pp. 156–174. (In Russian). EDN: SQVTKF.

2. Badmaeva, M. H. (2022) Etika iskusstvennogo intellekta: princip otvetstvennosti Gansa Jonasa [Ethics of Artificial Intelligence: Hans Jonas's Principle of Responsibility]. *Vestnik Buryatskogo gosudarstvennogo universiteta – The Buryat State University Bulletin*. No. 1. Pp. 67–79. (In Russian). DOI: 10.18101/1994-0866-2022-1-67-79. EDN: WAKAJS.

3. Belikova, E. K. (2024) Osnovnye voprosy filosofii iskusstvennogo intellekta [Key Issues in the Philosophy of Artificial Intelligence]. *Filosofiya i kul'tura – Philosophy and Culture*. No. 1. Pp. 1–11. (In Russian). DOI: 10.7256/2454-0757.2024.1.69543. EDN: PURYRC.

4. Razin, A. V. (2019) Etika iskusstvennogo intellekta [Ethics of Artificial Intelligence]. *Filosofiya i obshchestvo – Philosophy and Society*. No. 1. Pp. 57–73. (In Russian). EDN: YPCXWS.

5. Ignat'ev, V. I. (2019) I gryadet «drugoj» aktor: ctanovlenie tekhnosub"ekta v kontekste dvizheniya k tekhnologicheskoj singulyarnosti [And the “other” actor is coming: the formation of the technosubject in the context of the movement towards technological singularity]. *Sociologiya nauki i tekhnologij – Sociology of Science and Technology*. Vol. 10. No. 1. Pp. 64–78. (In Russian). DOI: 10.24411/2079-0910-2019-10005. EDN: XPCHXJ.

6. Karpov, V. E., Gotovcev, P. M, Rojzenzon, G. V. (2018) K voprosu ob etike i sistemah iskusstvennogo intellekta [On the issue of ethics and artificial intelligence systems]. *Filosofiya i obshchestvo – Philosophy and Society*. No. 2. Pp. 84–105. (In Russian). DOI: 10.30884/jfio/2018.02.07. EDN: YAEVYT.

7. Kolesnik, M. A., Kopceva, N. M. (2024) Filosofskie osnovy cifrovogo gumanizma [Philosophical Foundations of Digital Humanism]. *Cifrovizaciya – Digitalization*. Vol. 5. No. 1. Pp. 18–34. (In Russian). EDN: DOHPLW.

8. Moiseenko, M. V. (2018) Vyzovy sovremennosti: iskusstvennyj intellekt. Eticheskij aspekt [Challenges of Modernity: Artificial Intelligence. Ethical Aspect]. *Gumanitarnyj vestnik – Humanitarian Bulletin*. No. 9. – Stat'ya 3. (In Russian). DOI: 10.18698/2306-8477-2018-9-547. EDN: VKGOYT.

9. Nagl', L. (2022) Cifrovye tekhnologii: razmyshleniya o razlichii mezhdu instrumental'noj racional'nost'yu i prakticheskim razumom [Digital Technologies: Reflections on the Difference between Instrumental Rationality and Practical Reason]. *Kantovskij sbornik – Kantian Journal*. Vol. 41. No. 1. Pp. 60–88. (In Russian). DOI: 10.5922/0207-6918-2022-1-3. EDN: HSHBOF.

10. Nazarova, YU. V., Kashirin, A. YU (2023) Ekzistencial'nye riski iskusstvennogo intellekta v kontekste aksiologii transgumanizma: (po materialam rabot N. Bostroma) [Existential risks of artificial intelligence in the context of the axiology of transhumanism: (based on the works of N. Bostrom)]. *Obshchestvo: filosofiya, istoriya, kul'tura – Society: philosophy, history, culture*. No. 11. Pp. 40–45. (In Russian). DOI: 10.24158/fik.2023.11.5. EDN: OGBJBQ.

11. Sokolov, YU. I. (2019) Ekzistencial'nyj risk tekhnologicheskoj singulyarnosti. *Problemy analiza riska – Issues of risk analysis*. Vol. 16. No. 3. Pp. 62–77. (In Russian). DOI: 10.32686/1812-5220-2019-16-3-62-77. EDN: HDDELB.

12. Hokins, D. (2024) Supermozg: revolyuciya v ponimanii chelovecheskogo i iskusstvennogo intellekta [Superbrain: A Revolution in Understanding Human and Artificial Intelligence]. Mosow: Al'pina PRO. (In Russian).

13. Shatkin, M. A. (2023) Social'no-filosofskie aspekty razvitiya generativnogo iskusstvennogo intellekta [Social and philosophical aspects of the development of generative artificial intelligence]. *Izvestiya Saratovskogo universiteta. Novaya seriya. Seriya: Filosofiya. Psihologiya. Pedagogika – News of Saratov University. Ser. Philosophy, psychology, pedagogy*. Vol. 23, No. 4. Pp. 41–418. (In Russian). DOI: 10.18500/1819-7671-2023-23-4-414-418. EDN: PBDAAR.

14. Bostrom, N. (2002) Existential risks: analyzing human extinction scenarios and related hazards. Journal of Evolution and Technology. Vol. 9. – URL: http://jetpress.org/volume9/risks.html (accessed: 20.04.2025).

15. Coeckelbergh, M. (2024) What is digital humanism? A conceptual analysis and an argument for a more critical and political digital (post)humanism. Journal of Responsible Technology. Vol. 17. Pp. 100073. DOI: 10.1016/j.jrt.2023.100073

16. Nida-Rümelin, J., Weidenfeld, N. (2022) Digital humanism: for a humane transformation of democracy, economy and culture in the digital age. Cham: Springer.

17. Floridi, L. (2023) AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models. Philosophy & Technology. Vol. 36. Article number 15. DOI: 10.1007/s13347-023-00621-y.

18. Jonas, H. (1982) Technology as a subject for ethics. Social Research. Vol. 49, No. 4. Pp. 891–898.

19. Nowotny, H. (2022) Digital humanism: navigating the tensions ahead. Perspectives on digital humanism. ed. H. Werthner, E. Prem, E. F. Lee, C. Ghezzi. Cham: Springer. Pp. 317–321

20. Prem, E. (2024) Principles of digital humanism: a critical post-humanist view. Journal of Responsible Technology. Vol. 17. Pp. 100075. DOI: 10.1016/j.jrt.2024.100075.

21. Russell, S. (2022) Artificial intelligence and the problem of control. Perspectives on digital humanism / ed. H. Werthner, E. Prem, E. F. Lee, C. Ghezzi. – Cham: Springer. Pp. 19–24.

22. Russell, S. (2019) Human compatible: artificial intelligence and the problem of control. London: Penguin.

Информация об авторе

Белобрагина Анна Сергеевна – аспирант, Национальный исследовательский университет «Московский институт электронной техники», Москва, Российская Федерация; ORCID ID: 0009-0001-0110-2223, e-mail: belobragina@gmail.com

Information about the author

Anna S. Belobragina – postgraduate student, National Research University of Electronic Technology, Moscow, Russian Federation; ORCID ID: 0009-0001-0110-2223, e-mail: belobragina@gmail.com

Поступила в редакцию: 28.05.2025
Принята к публикации: 18.06.2025
Опубликована: 25.09.2025

Received: 28 May 2025
Accepted: 18 June 2025
Published: 25 September 2025

ГРНТИ: 02.31.31

ВАК: 5.7.6