Экономические и социально-гуманитарные исследования. 2025. Т. 12. № 2. С. 137–145.
Economic and Social Research. 2025. Vol. 12. No. 2. P. 137–145.
Научная статья

УДК 111.1 + 165.0 + 17.03 + 004.8
DOI: 10.24151/2409-1073-2025-12-2-137-145
EDN: QVIFKL

Искусственный интеллект: о «новой этике» М. Габриэля

И. Е. Прись

Институт философии Национальной академии наук Беларуси, г. Минск, Беларусь

frigpr@gmail.com

Аннотация. Критически оценивается «новая этика» искусственного интеллекта, предложенная М. Габриэлем. Утверждается, что в отличие от интеллекта человека искусственный интеллект лишен нормативного измерения или, что эквивалентно, чувствительности к контексту. Автор показывает противоречие между точкой зрения М. Габриэля и реалистическим контекстуальным подходом к этике у Ж. Бенуа, и моральным реализмом Т. Уильямсона, согласно которым первичны не принципы, а моральное восприятие в контексте, парадигматические примеры морального знания. Сравниваются подходы к пониманию искусственного интеллекта М. Габриэля, Д. Андлера, Л. Флориди, С. Рассела. Доказывается целесообразность принципа умеренности Д. Андлера. Реалистическая концепция искусственного интеллекта (ИИ) противопоставляется идеалистической концепции.

Ключевые слова: искусственный интеллект, этика ИИ, Габриэль, моральный прогресс, автономия, контекст, нормативность, моральный реализм, принцип умеренности

Финансирование работы: работа выполнена в рамках НИР «Сознание и искусственный интеллект в условиях цифровых трансформаций: научно-методологический и социогуманитарный аспекты» Института философии НАН Беларуси.

Для цитирования: Прись И. Е. «Искусственный интеллект: о „новой этике“ М. Габриэля». *Экономические и социально-гуманитарные исследования* 12.2 (2025): 137–145. https://doi.org/10.24151/2409-1073-2025-12-2-137-145 EDN: QVIFKL.

Original article

Artificial Intelligence: On M. Gabriel's "New Ethics"

I. E. Pris

Institute of Philosophy of the National Academy of Sciences of Belarus, Minsk, Belarus

frigpr@gmail.com

Abstract. The "new ethics" of artificial intelligence proposed by M. Gabriel is critically evaluated. It is argued that, unlike human intelligence, artificial intelligence is devoid of a normative dimension, or, equivalently, context sensitivity. The author shows the contradiction between M. Gabriel's viewpoint and J. Benoist's contextual realist approach to ethics, and T. Williamson's moral realism, according to which it is not principles that are primary, but moral perception in context, paradigmatic examples of moral knowledge. The approaches to artificial intelligence by M. Gabriel, D. Andler, L. Floridi, and S. Russell are compared. The validity of D. Andler's principle of moderation is affirmed. The realist conception of artificial intelligence (AI) is opposed to its idealist conception.

Keywords: artificial intelligence, AI ethics, Gabriel, moral progress, autonomy, context, normativity, moral realism, moderation principle

Funding: the work has been carried out within the project "Consciousness and Artificial Intelligence in the conditions of digital transformations: Scientific, methodological and socio-humanitarian aspects" of Institute of Philosophy of NAS of Belarus.

For citation: Pris I. E. "Artificial Intelligence: On M. Gabriel's 'New Ethics' ". *Ekonomicheskie i sotsial'no-gumanitarnye issledovaniya = Economic and Social Research* 12.2 (2025): 137–145. (In Russian). https://doi.org/10.24151/2409-1073-2025-12-2-137-145

Введение

Быстрое развитие систем искусственного интеллекта (ИИ) влечет за собой возникновение новых этических проблем. Выделим три взаимосвязанных значения термина «этика искусственного интеллекта»[1]. Во-первых, это исследование этических вопросов, связанных с производством и применением ИИ. Во-вторых, исследование возможностей создания ИИ этического согласно дизайну. Наиболее общие принципы здесь те же, что и в медицинской этике — благодеяние, непричинение вреда, справедливость, автономия (человека), плюс специфический для ИИ принцип объяснямости. Принципы могут незначительно варьироваться (Andler, 2023; Coeckelbergh, 2020). Так, П. Брей и Б. Дейноу упоминают шесть принципов: «свобода (они еще говорят о человеческой агентности, включающей в себя свободу, автономию и достоинство. — *И. П.*), неприкосновенность частной жизни (privacy), справедливость, прозрачность, подотчетность и благополучие (отдельных людей, общества и окружающей среды)» (Brey, Dainow, 2024: 4: 1267—1268). К ним можно добавить безвредность и ответственность. Эти абстрактные принципы

[1] Специалисты включают в трактовку понятия «искусственный интеллект» не только программы, алгоритмы, запрограммированные компьютеры и роботы (системы ИИ), но также и соответствующие лаборатории, институты, проекты. Обычно по контексту употребления термина ИИ можно понять, о чем идет речь. В будущем, возможно, этот термин будет также обозначать некоторое новое общее свойство, присущее всем системам ИИ.

дополняются более оперативными. Наконец, в-третьих, возникает вопрос об ИИ, который обладал бы способностью открывать или производить новые этические ценности.

Цель новой этики ИИ, или этики Нового Просвещения, предлагаемой немецким философом М. Габриэлем, можно понимать так: в процессе глобальной кооперации культур, обладающих различными ценностями, создать мощный этический по своему дизайну ИИ, некоего АльфаБудду, или АльфаИисуса, который бы открывал или помогал человеку открывать и социально-экономически воплощать новые моральные факты и законы (в том числе и в отношении самого ИИ), т. е. активно способствовал бы не просто радикальным изменениям общества, а рационально контролируемому, научно направляемому моральному прогрессу. Такой ИИ рассматривается Габриэлем как система универсализации морали, помогающая понять, кто мы есть как человеческие существа, кем мы хотим быть и кем мы должны стать[2]. Этот проект вызывает некоторые оговорки и опасения — в частности, они касаются возможной потери человеком его автономии, по крайней мере частично.

ИИ vs естественный интеллект

Прежде всего оценка проекта Габриэля требует ответа на вопрос «как соотносятся ИИ и интеллект человека?». Мы трактуем отношение между ними в терминах категориального различия между идеальным (нормативным) и реальным. Это различие можно также объяснить в терминах витгенштейновской проблемы следования правилу. ИИ следует формальным (машинным) правилам (Прись, 2024a; Прись, 2024b). Близкая точка зрения отстаивалась С. Шэнкером в его книге «Замечания Витгенштейна об основаниях ИИ» еще в 1998 г. (Shanker, 1998). Это также согласуется с тем, что ИИ, по мнению французского философа Д. Андлера, относится к человеку как тень к ковбою, а по мнению Габриэля — как карта (ИИ) к территории (человек)[3] (Andler, 2023). Для Габриэля ИИ — модель мысли, так как ИИ имеет не нейробиологическое, а искусственное основание (Gabriel, 2018). Расхождение между нами и Габриэлем — в следующем: для Габриэля мысль реальна, является чем-то вроде неприродного шестого чувства человека[4], в отличие от не имеющего собственной реальности информационного процесса (с этой точки зрения ИИ не мыслит), тогда как для нас она идеальна, но это подразумевает ее укорененность в реальности, в том числе и нейробиологической (согласно концептуальной грамматике понятия «мысль»).

На этапе предыдущего исследования мы утверждали, что ИИ не интеллект и в рамках существующей натуралистической парадигмы никогда им не будет, поскольку он лишен нормативного измерения, или, что эквивалентно, чувствительности к контексту. Идея трансгуманизма — миф. Так называемый момент сингулярности никогда не наступит (Прись, 2024a; Прись, 2024b)[5]. В то же время прометеевский проект создания автономного ИИ по образу и подобию человека несет в себе угрозу, поэтому от него следует отказаться. Схожую позицию занимает

[2] Gabriel Markus, Scheu René. “The New Ethics of AI”. DLD 24. *YouTube.* 11 Jan. 2024. Web. 30 Jan. 2025. <https://www.youtube.com/watch?v=ynBU7untOoc>.

[3] Ibid.

[4] По этой причине для Габриэля человеческий интеллект — «искусственный интеллект» (но, разумеется, не в том смысле, в котором мы говорим об ИИ) (Gabriel, 2018).

[5] Из современных философов этой же точки зрения придерживаются, например, М. Габриэль, Д. Андлер, Л. Флориди, М. Битболь. Противоположной точки зрения придерживается, например, Д. Чалмерс (Chalmers, 2010).

[6] «Интеллект — это не вещь, не явление, не процесс и не функция, а норма, которая применяется к поведению: он квалифицирует отношение между человеком и его миром, причем таким образом, который никогда не является объективным и окончательным» (Andler, 2023: 12).

Д. Андлер: контекст имеет нормативное измерение, а интеллект — это нормативность[6], тогда как ИИ лишь способен решать проблемы, что является второстепенной задачей для человеческого интеллекта (Andler, 2023; Andler, 2000). Маркус Габриэль, напротив, определяет интеллект как способность решать проблемы. В этом смысле ИИ может быть умнее, чем человек, хотя и не обладает высшей формой мышления — рефлексивной. Также в работах Габриэля можно найти высказывание, что никто не знает, что такое мышление, мысль. Если мысль есть нечто абстрактное, процесс в реальности, необязательно предполагающий нейробиологическое основание, то даже о современных системах ИИ можно сказать, что они мыслят. В этом случае, согласно Габриэлю, модель, т. е. ИИ, принадлежала бы той же «реальности», что и моделируемая система, т. е. мысль человека.

Итальянский философ Л. Флориди считает, что вопрос о том, мыслит ИИ или нет, не имеет значения (Floridi, 2023). Важно то, что ИИ делает и способен делать. Флориди считает, что ИИ не мыслит, но является агентом. ИИ — новый вид агентства. Это нечеловеческое, бездумное агентство, преобразующее окружающую среду и требующее ее преобразования (семантизации, наполнения содержанием). В противном случае ИИ не мог бы использоваться. Но если под агентством понимать способность совершать полноценные действия, нельзя назвать искусственные системы агентами. Действия, как и суждения, нормативны. К ним способны только люди.

Новый моральный реализм vs контекстуальный реализм

Согласно новому моральному реализму Габриэля, существуют универсальные, априорные, абсолютные и неизменные моральные принципы, которые сначала открываются, а затем применяются во внешнем по отношению к ним контексте (Gabriel, 2020). Этот неоклассический подход к морали вступает в противоречие с реалистическим контекстуальным подходом французского философа Ж. Бенуа, который мы разделяем, и моральным реализмом британского философа Т. Уильямсона, критикующего моральный инференциализм (Прись, 2023; Williamson, 2022). Проблематична и более общая позиция — моральный принципиализм (различные принципы могут противоречить друг другу, по-разному интерпретироваться, их применимость зависит от контекста). На самом деле первичны не принципы, а моральное восприятие в контексте, парадигматические примеры морального знания (Прись, 2023).

Следует также принять во внимание уильямсоновскую критику интернализма и когерентизма в эпистемологии, а также витгенштейновскую критику понятия морального факта, в котором содержались бы все его применения. Этика не может обойтись без онтологии (моральных фактов), но и не может редуцироваться к онтологии. То, что есть, не может сказать о том, что должно быть. Другими словами, введение платонизирующего (идеального), но неметафизического измерения необходимо (Прись, 2023). Но это как раз то, чего ИИ лишен по определению.

Новая этика ИИ, на наш взгляд, подразумевает общий подход Габриэля к морали (Gabriel, 2020). Но если ИИ не чувствителен к контексту (в противном случае он не был бы ИИ, а был бы человеком или, быть может, неким автономным нечеловеческим интеллектом с нечеловеческой моралью), тем более контексту моральному — а суть морали заключена в такой чувствительности, — то возникает вопрос о возможности реализации предлагаемой Габриэлем программы морального прогресса при помощи ИИ и о потенциальных последствиях попыток ее реализации. Возможно, моральный проект

Габриэля «Будь прогрессивным!» следует заменить более умеренным проектом.

На пути к кантовскому и витгенштейновскому ИИ

Классический символический ИИ (1) — это программа, алгоритм, расширенная логика. Сменивший его коннекционистский ИИ (2) — искусственная нейронная сеть. Философия первого — рационализм (всё — логика!); философия второго — эмпиризм (всё — перцепция!), хотя он содержит в себе и существенные элементы символического ИИ. Так называемое глубокое обучение и большие языковые модели (ChatGPT и др.) — это современное развитие коннекционизма. Предположительно, ИИ в ближайшем будущем синтезирует оба подхода. Философию такого гибридного ИИ мы бы условно сравнили с кантовским критическим синтезом рационализма и эмпиризма[7].

Соответственно, этика может быть встроена в ИИ сверху вниз (кажется, этот подход ближе принципиалистскому подходу Габриэля), но она может быть встроена в него и снизу вверх, путем обучения ИИ на больших массивах эмпирических данных.

Так, С. Рассел предлагает альтернативу принципиализму. Суть его подхода в том, чтобы ориентировать этику ИИ на предпочтения человека, которые бы выявлялись из статистических данных о его поведении (Russell, 2020: ch. 7). Этот подход — индуктивизм, как отмечает Д. Андлер, — основан на иллюзиях. На самом деле невозможно чисто статистически выявить предпочтения человека, поведение определяется далеко не только предпочтениями; и, наконец, будущее не всегда должно определяться прошлым — как такое будущее, которое имеет высокую вероятность наступления (в кризисных и неразрешимых ситуациях, а также в науке и искусстве такая вероятностная рациональность непригодна) (Andler, 2023: 223).

ИИ как новый вид реальности

ИИ — новый вид реальности. Но он не существует сам по себе (абсолютно), а интегрирован в социально-экономические, культурные, политические, материальные и другие отношения, практики, т. е. имеет реальные условия своего существования. Если мы перестанем заботиться о нем, он исчезнет. ИИ — сложная технология. Как известно, когда сложная технология используется большим количеством независимых агентов, возникают ситуации, когда не агенты контролируют технологию, а технология агентов, что говорит о ее реальности.

Существует общая проблема контроля ИИ, в частности проблема согласования

[7] Аналогичное сравнение делает Р. Эванс. Он пишет: «Нейронная сеть — интеллектуальный предок (ancestor) эмпиризма, так же как логическое (logic-based) обучение — интеллектуальный предок рационализма. Кантовское объединение эмпиризма и рационализма — когнитивная архитектура, которая пытается комбинировать лучшее из обоих миров и указывает путь к гибридной архитектуре, комбинирующей лучшее из нейронных сетей и логических подходов» (Evans, 2022: 41).

Ряд ученых считает, что кантовский категорический императив может быть формализован, алгоритмизирован и встроен в ИИ (см., напр: Evans, 2022; Powers, 2006; Lindner, Bentzen, 2019). Другие заключают, что ИИ не может быть подлинным кантовским моральным агентом, поскольку он не обладает свободой воли, автономией и рациональностью в кантовском смысле (Chakraborty, Bhuyan, 2024). С точки зрения нашего контекстуального / нормативного подхода, последнее заключение очевидно. Вместе с тем ИИ, имитирующий этического агента, возможен и имеет практическое значение. Например, в статье З. Гудмунсена утверждается, что ИИ может реагировать на рациональные, в частности моральные, основания, выносить (моральные) суждения и принимать (моральные) решения. Вместе с тем в той же статье утверждается, что ИИ не может быть аутентичным, или ответственным (моральным) агентом (Gudmunsen, 2025). Соглашаясь лишь с последним, отметим, что (аутентичные) восприимчивость к рациональным основаниям, а также суждения и решения являются нормативными, тогда как у ИИ они чисто каузальны и, следовательно, представляют собой лишь имитацию.

[8] В русскоязычной литературе употребляются еще термины «проблема сонастройки» и «проблема выравнивания».

(alignment) этики ИИ с этикой человека[8]. Мы не в состоянии полностью контролировать ИИ. Поэтому мы хотим, чтобы по крайней мере ценности, вложенные в ИИ, совпадали или согласовывались с ценностями человека. Эта проблема может оказаться неразрешимой (Andler, 2023: § 10.5)[9]. Дилемма здесь следующая: либо мы конструируем такие ИИ, которые не могут решать сложные задачи, которые мы без помощи ИИ тоже не можем решать, но хотели бы, чтобы они были решены; либо мы конструируем такие ИИ, которые способны решать сложные задачи, но в то же время оказываются, по крайней мере частично, (квази)автономными[10]. Проблема в том, что (квази)автономной системе невозможно навязать ценности извне по определению. Она сама выбирает ценности и выбирает, принять или нет ту или иную предлагаемую ей систему ценностей.

Аспектом проблемы согласования является проблема определения того, какие человеческие ценности должны быть приоритетными для согласования, чьи ценности должны быть закодированы в системах ИИ. Это проблема «ценностного плюрализма, при котором различные люди и культуры придерживаются разнообразных, противоречивых и несводимых друг к другу ценностей. Недемократическое согласование ценностей лишает пользователей возможности выступать в качестве полноценных эпистемических агентов и, как следствие, полноценных моральных агентов» (Huang, Papyshev, Wong, 2025: 14). Трудно, а быть может, невозможно сделать так, чтобы ИИ одновременно принял во внимание интересы общества в целом, различных групп людей и отдельных людей[11]. Также существуют разные нормативные этические теории. Мысленный эксперимент с автономным автомобилем как версия классического мысленного эксперимента «проблема вагонетки» иллюстрирует проблему. В зависимости от заложенной в программу ИИ системы нормативной этики — деонтологической или утилитаристской, а также от их интерпретаций, ИИ будет «поступать» так или иначе в некоторых хорошо определенных (соответствующих алгоритму ИИ) ситуациях (см. анализ проблемы, в частности в кантианской перспективе: (Huang, Papyshev, Wong, 2025; Kim, Schönecker, eds, 2022: ch. 6—8)).

Принцип умеренности

Но даже если ИИ будет относительно безопасен, мы можем попасть в зависимость от него, поскольку, живя в преобразованном под него мире, мы уже больше не сможем без него обойтись. Возникает вопрос: хотим ли мы жить в мире, созданном для машин, и не иметь возможности обойтись без них?

Андлер, например, выдвигает принцип умеренности, следование которому

[9] В научной дискуссии обсуждается также связанная с проблемой согласования «проблема разрыва ответственности» (responsibility gap), которая поднимает вопрос о том, кто несет ответственность за непредсказуемые действия, совершаемые самообучающимся (квази)автономным ИИ. На наш взгляд, попытка переложить ответственность, хотя бы частично, на ИИ — несостоятельна.

[10] Сегодня ИИ считается автономным, если его поведение не детерминировано исходным, вложенным в него знанием, а он самостоятельно способен интегрировать новую эмпирическую информацию из окружающей среды и вести себя в соответствии с ней, вообще говоря, непредсказуемым образом. (Если бы ИИ был полностью предсказуем, мы бы не нуждались в его услугах.) Таким образом, это не кантовское понимание автономии.

[11] Для решения проблемы согласования применяется философия позднего Витгенштейна (Perez-Escobar, Sarikaya, 2024). Предлагается принять во внимание психологический, социальный и культурный контексты, их изменчивость. Несмотря на то что такой подход действительно позволяет снизить остроту проблемы, он, как представляется, основан на имитации чувствительности к контексту. Никакого подлинного следования правилу в том смысле, в котором его понимает Витгенштейн, здесь нет. Что же касается имитации витгенштейновского ИИ, то она возможна, но сложнее, чем имитация кантовского ИИ (см. попытки использовать ресурсы философии Канта для улучшения «когнитивных» и «этических» способностей ИИ: (Evans, 2022; McDonald, 2023; Kim, Schönecker, eds, 2022; Schlicht, 2022)).

представляется наиболее целесообразным: «Используйте искусственный интеллект только тогда, когда риски снижены, а выгоды значительны; используйте системы ИИ, которые максимально просты и способны обеспечить ожидаемый сервис» (Andler, 2023: 224). Этот принцип предполагает: 1) не поручать ИИ те задачи, которые можно решить без него; 2) не придавать ИИ гуманоидное обличье; 3) не использовать ИИ там, где требуется интеллект человека (т. е. не только способность решать проблемы), — в частности, не поручать ИИ задачи, решение которых требует мудрости.

Квантовая логика в известном смысле принимает во внимание не(пред)определенность и контекстуальность, свойственные решениям и поступкам людей. Можно поэтому предположить, что основанный на ней квантовый или квантовоподобный ИИ будет человекоподобным (Прись, 2024c). Но и такой ИИ никогда не станет умным и этическим, не приблизится к человеку, так как контекст не редуцируется к логическим операциям.

Заключение

ИИ — имитация интеллекта, этики, автономии, агентства (действия)[12]. Концептуальные смешения искусственного и естественного, идеального и реального имеют нежелательные последствия, как теоретические, так и практические. Одна из задач философии ИИ как раз в том и состоит, чтобы отделить одно от другого, максимально выявить различия между ИИ и человеком. Всё, что способен или будет способен делать ИИ, каким бы совершенным он ни был, не относится к природе человека. Другими словами, нам нужна реалистическая, а не идеалистическая концепция ИИ.

Список литературы и источников / References

Прись И. Е. «Искусственный интеллект и неоэкзистенциализм». *Философия в XXI веке: направления и тенденции развития: материалы II Междунар. науч.-практ. конф. (Москва, Зеленоград — Красноярск, 12 апреля 2024): в 3 ч.* Под общ. ред. Н. В. Даниелян. Ч. 2. М.: МИЭТ, 2024a. 159—169. EDN: FOARAA.

Pris I. E. “Artificial Intelligence and Neo-Existentialism”. *Filosofiya v XXI veke: napravleniya i tendentsii razvitiya: materialy II Mezhdunar. nauch.-prakt. konf. (Moskva, Zelenograd — Krasnoyarsk, 12 aprelya 2024).* Gen. ed. N. V. Danielyan. Moscow: MIET, 2024a. 159—169. (In Russian). 3 parts.

Прись И. Е. «Искусственный интеллект — не интеллект и никогда им не будет». *Наука и инновации* 9 (259) (2024b): 26—29. EDN: CDNTHD.

Pris I E. “Artificial Intelligence Is Not Intelligence and Never Will Be”. *Nauka i innovatsii = Science and Innovations* 9 (259) (2024b): 26—29. (In Russian).

Прись И. Е. «Квантовоподобное моделирование и его философские основания». *Философия науки* 3 (102) (2024c): 109—129. https://doi.org/10.15372/PS20240307. EDN: VJUFIK.

Pris I. E. “Quantum-Like Modeling and its Philosophical Foundations”. *Filosofiya nauki = Philosophy of Science* 3 (102) (2024c): 109—129. (In Russian). https://doi.org/10.15372/PS20240307

[12] Могут сказать: «Но ведь это очевидно!» И, с нашей точки зрения, это действительно так. Философское исследование ИИ не столько доказывает отсутствие у ИИ подлинного интеллекта, этики, сколько пытается выявить то, что к ИИ не относится, т. е. природу естественного интеллекта, человека. Вопрос «что такое человек?» Кант, как известно, считал ключевым вопросом философии.

В то же время непредсказуемость ИИ не позволяет считать, что ИИ есть только лишь имитация естественного интеллекта. Системы ИИ также можно рассматривать как новый вид реальности, для которой традиционные понятия обобщаются или приобретают другой смысл. Например, можно ввести неантропоморфное понятие ИИ, заслуживающего доверия (Simion, Kelp, 2023). Также были предложены и обоснованы такие понятия, как «искусственный агент», «искусственный эпистемический авторитет», «обобщенная (применимая к ИИ) пропозициональная установка» и др.

Прись И. Е. «Контекстуальный моральный реализм». *Сибирский философский журнал* 21.4 (2023): 5—28. https://doi.org/10.25295/2541-7517-2023-21-4-5-28. EDN: MIEJMS.
Pris I. E. "Contextual Moral Realism". *Sibirskij filosofskij žurnal = Siberian Journal of Philosophy* 21.4 (2023): 5—28. (In Russian). https://doi.org/10.25295/2541-7517-2023-21-4-5-28

Andler D. *Intelligence artificielle, intelligence humaine: la double énigme*. Paris: Gallimard, 2023. 432 p.

Andler D. "The Normativity of Context". *Philosophical Studies* 100.3 (2000): 273—303. https://doi.org/10.1023/A:1018628709589

Brey P., Dainow B. "Ethics by Design for Artificial Intelligence". *AI and Ethics* 4.4 (2024): 1265—1277. https://doi.org/10.1007/s43681-023-00330-4

Chakraborty A., Bhuyan N. "Can Artificial Intelligence Be a Kantian Moral Agent? On Moral Autonomy of AI System". *AI and Ethics* 4 (2024): 325—331. https://doi.org/10.1007/s43681-023-00269-6

Chalmers D. "The Singularity: A Philosophical Analysis". *Journal of Consciousness Studies* 17.9-10 (2010): 7—65.

Coeckelbergh M. *AI Ethics.* Cambridge, MA: The MIT Press, 2020. 248 p.

Evans R. "2 The Apperception Engine". Kim H., Schönecker D., eds. *Kant and Artificial Intelligence*. Berlin: De Gruyter, 2022. 39—103. https://doi.org/10.1515/9783110706611-002

Floridi L. *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities.* Oxford: Oxford Up, 2023. 272 p.

Gabriel M. *Der Sinn des Denkens.* Berlin: Ullstein, 2018. 368 S.

Gabriel M. *Moralischer Fortschritt in dunklen Zeiten: Universale Werte für das 21. Jahrhundert.* Berlin: Ullstein, 2020. 369 S.

Gudmunsen Z. "The Moral Decision Machine: A Challenge for Artificial Moral Agency Based on Moral Deference". *AI and Ethics* 5 (2025): 1033—1045. https://doi.org/10.1007/s43681-024-00444-3

Huang L. T.-L., Papyshev G., Wong J. K. "Democratizing Value Alignment: From Authoritarian to Democratic". *AI and Ethics* 5 (2025): 11—18. https://doi.org/10.1007/s43681-024-00624-1

Kim H., Schönecker D., eds. *Kant and Artificial Intelligence*. Berlin: De Gruyter, 2022. v, 290 p. https://doi.org/10.1515/9783110706611

Lindner F., Bentzen M. M. "A Formalization of Kant's Second Formulation of the Categorical Imperative". *arXiv.* Rev. 11 July 2019. Web. 15 May 2025. https://doi.org/10.48550/arXiv.1801.03160

McDonald F. J. "AI, Alignment, and the Categorical Imperative". *AI and Ethics* 3 (2023): 337—344. https://doi.org/10.1007/s43681-022-00160-w

Perez-Escobar J. A., Sarikaya D. "Philosophical Investigations into AI Alignment: A Wittgensteinian Framework". *Philosophy & Technology* 37.3 (2024): 80. https://doi.org/10.1007/s13347-024-00761-9

Powers T. M. "Prospects for a Kantian Machine". *IEEE Intelligent System* 21.4 (2006): 46—51. https://doi.org/10.1109/MIS.2006.77

Russell S. *Human Compatible. Artificial Intelligence and the Problem of Control.* New York: Penguin Books, 2020. 352 p.

Schlicht T. "1 Minds, Brains, and Deep Learning: The Development of Cognitive Science through the Lens of Kant's Approach to Cognition". Kim H., Schönecker D., eds. *Kant and Artificial Intelligence*. Berlin: De Gruyter, 2022. 3—38. https://doi.org/10.1515/9783110706611-001

Shanker S. G. *Wittgenstein's Remarks on the Foundations of AI.* London: Routledge, 1998. xvi, 280 p.

Simion M., Kelp Ch. "Trustworthy Artificial Intelligence". *Asian Journal of Philosophy* 2 (2023): 8. https://doi.org/10.1007/s44204-023-00063-5

Williamson T. "Unexceptional Moral Knowledge". *Journal of Chinese Philosophy* 49.4 (2022): 405—415. https://doi.org/10.1163/15406253-12340082

Информация об авторе

Прись Игорь Евгеньевич — доктор философии (PhD), кандидат физико-математических наук, ведущий научный сотрудник Института философии Национальной академии наук Беларуси (Беларусь, 220072, г. Минск, ул. Сурганова, 1, корп. 2), *frigpr@gmail.com*, SPIN-код: 2663-6744, ORCID: 0000-0003-1721-6388.

Information about the author

Igor E. Pris — Doctor in Philosophy (PhD), Cand. Sci. (Theor. Phys.), Leading Researcher, Institute of Philosophy of the National Academy of Sciences of Belarus (Belarus, 220072, Minsk, Surganova st., 1/2), *frigpr@gmail.com*, SPIN code: 2663-6744, ORCID: 0000-0003-1721-6388.

Статья поступила в редакцию 10.12.2024, одобрена после рецензирования 06.02.2025.
The article was submitted 10.12.2024, approved after reviewing 06.02.2025.