Дискурсы этики. 2024, 3–4(23–24): 13—24
ISSN 2311-570X (online)
Постоянная ссылка:
http://theoreticalappliedethics.org/wp-content/uploads/2025/02/DE2024_3-4_23-24_13-24.pdf

УДК 17; 123

ИДЕЯ МОРАЛЬНОЙ АВТОНОМИИ В ФИЛОСОФИИ И. КАНТА И ЭТИЧЕСКИЕ ПРОБЛЕМЫ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Перов В. Ю.

статья:
поступила в редакцию 25.11.2024
принята к публикации 28.12.2024
опубликована (онлайн) 28.02.2025

© Перов Вадим Юрьевич
кандидат философских наук, доцент, заведующий каф. этики Института философии Санкт-Петербургского государственного университета, Санкт-Петербург, Россия
адрес для корреспонденции: vadimperov@gmail.com

Аннотация. Статья посвящена моральной автономии, которая является одной из основных идей в современной этической мысли, связанной с философией И. Канта. Моральная автономия означает, что моральные принципы не должны зависеть от внешних обстоятельств или чужого влияния. Согласно Канту, основополагающее значение имеет чистая нравственность, которая вытекает из деятельности чистого практического разума. Поступки, обусловленные внешними факторами, рассматриваются как проявление гетерономии воли, в то время как моральный закон присущ только разумным существам. Кант считал, что человек как разумное существо обладает способностью самому создавать моральные законы и свободно следовать им без влияния целей счастья, поль-

зы или удовольствия. Эта идея имеет значительное значение в современных этических исследованиях и обсуждениях, особенно в контексте развития систем искусственного интеллекта (ИИ), которые обладают аналогом разумности. Существует точка зрения, что технологии ИИ могли бы претендовать на цифровой аналог «чистого практического разума», поскольку лишены всех, свойственных людям недостаткам, мешающим принимать морально правильные решения (свобода (автономия) от эмоций и независимость от внешнего влияния и т.д.). Вопросы о возможности существования искусственных моральных агентов (ИМА) и их влиянии на моральную автономию людей становятся все более актуальными. В ходе исследования выделены следующие основные группы проблем. Во-первых, возможности и границы моральной автономии искусственного интеллекта. Одной из пока неразрешимых этических проблем, связанных с созданием ИИ, является его «моральная предвзятость», которая может быть частично устранена только людьми. Таким образом, речь может идти только об их функциональной автономии, то есть о способности работать какое-то время без непосредственного вмешательства со стороны людей в соответствии с заранее определенным набором целей, даже если способы их достижения частично формируются в ходе работы ИИ. Во-вторых, потенциальные влияния и риски для моральной автономии людей при взаимодействии с искусственным интеллектом, что наиболее четко проявилось в возникновении о «смерти приватности». В-третьих, ещё большие этические риски связаны с возможностью моральной субституции, то есть перекладывания решения моральных вопросов на алгоритмы ИИ, даже в тех случаях, когда последние выступают в качестве источников морально значимой информации. В качестве главного вывода выдвигается идея о том, что нравственность в силу специфики моральной автономии по-прежнему остается делом людей.

Ключевые слова: автономия моральная, практический разум, искусственный интеллект, И. Кант, моральная предвзятость, этические риски.

Исследование проведено в рамках проекта РНФ №24-28-00562 «Философские основания этических рисков в сфере систем искусственного интеллекта».

Discourses of Ethics. 2024, 3–4(23–24): 13—24
ISSN 2311-570X (online)
permanent link:
http://theoreticalappliedethics.org/wp-content/uploads/2025/02/DE2024_3-4_23-24_13-24.pdf

THE IDEA OF MORAL AUTONOMY IN THE PHILOSOPHY OF I. KANT AND ETHICAL PROBLEMS OF ARTIFICIAL INTELLIGENCE SYSTEMS

Vadim Perov

received 25.11.2024
accepted 28.12.2024
published (online) 28.02.2025

Candidate of Science in Philosophy, Docent, Head of the Department of Ethics,
Institute of Philosophy, Saint Petersburg State University, St. Petersburg, Russia
Correspondence to: vadimperov@gmail.com

Abstract. The article is devoted to moral autonomy, which is one of the main ideas in modern ethical thought associated with the philosophy of I. Kant. Moral autonomy means that moral principles should not depend on external circumstances or the influence of other people. This idea is importance in contemporary ethical research and discussions, especially in the context of the development of artificial intelligence (AI) systems, which could claim to be the digital analogue of "pure practical reason". Questions about the possibility of the existence of artificial moral agents (AMA) and their impact on the moral autonomy of people are becoming increasingly relevant. The study identifies the following main groups of problems. First, the possibilities and limits of the moral autonomy of artificial intelligence. One of the so far unsolvable ethical problems associated with the creation of AI is its "moral bias", which can only be partially eliminated by humans. Thus, we can only talk about their functional autonomy, that is, the ability to work for some time without direct human

intervention in accordance with a predetermined set of goals, even if the ways to achieve them are partially formed during the work of AI. Secondly, the potential consequences and risks for people's moral autonomy when interacting with artificial intelligence, which were most clearly manifested in the emergence of the "death of private life". Thirdly, even greater ethical risks are associated with the possibility of moral substitution, that is, shifting the solution of moral issues to artificial intelligence algorithms, even in cases where the latter act as sources of morally significant information. As the main conclusion, the idea is put forward that morality, due to the specifics of moral autonomy, is still a matter for people.

Keywords: Moral Autonomy, Practical Reason, Artificial Intelligence, I. Kant, Moral Bias, Ethical Risks.

Funded by Russian Science Foundation (RSF) №24-28-00562 "Philosophical foundations of ethical risks in the field of artificial intelligence systems".

Одним из фундаментальных вкладов в этику, который традиционно связывают с творчеством И. Канта — это обоснование идеи о принципиальной автономии морали (моральной автономии). Хотя следует признать, что некоторые аналоги моральной автономии в смысле «самозакония» можно найти в истории этической мысли. Для этого достаточно упомянуть автаркию у стоиков и всю традицию отстаивания необходимости и безусловной нравственной ценности свободных и самостоятельных, то есть независимых от внешних обстоятельств и чужого влияния, моральных суждений и оценок. При этом есть достаточные основания полагать, что именно в этике Канта идея моральной автономии приобретает тот вид и значение, которое в современных теоретической и прикладных этиках является предметом многочисленных исследований и обсуждений в различных контекстах. Не имея возможности проанализировать имеющиеся многочисленные интерпретации кантовского понимания моральной автономии, имеет смысл кратко остановиться на ключевых, ставших почти хрестоматийными, ее характеристиках.

1) В центре внимания Канта было то, что может быть названо «чистой нравственностью» как результат деятельности «чистого практического разума». В какой-то степени это можно рассматривать как своеобразное продолжение теоретико-методологической установки философии Нового времени на «очищение разума» от всех эмоционально-психологических, социальных, культурных и т.д. наслоений, то есть внешних по отношению к самому разуму обстоятельств.

2) Поступки, обусловленные этими внешними факторами (счастье, польза, удовольствие и т.д.), рассматривались Кантом как проявление гетерономии воли. Такая воля является несвободной, преследуемые в таких поступках цели случайны, правила, по которым они осуществляются, имеют форму гипотетических императивов и не могут быть моральным законом для всех разумных существ. Гетерономные поступки

в лучшем случае являются «советами благоразумия», но в этом смысле не имеют отношения к нравственности, поскольку, строго говоря, являются внеморальными.

3) Чистый практический разум сам себе дает моральный закон (по сути, он и есть этот моральный закон), который в силу дуалистичной феноменально-ноуменальной природы человека принимает форму категорического императива, обязательного для всех.

4) Именно эта способность к моральному самозаконодательству и свободному подчинению ему на основе чувства долга по отношению к категорическому императиву исключительно в силу разумности последнего являются, согласно Канту, основанием для признания морального достоинства людей как разумных существ.

Последнее обстоятельство в этике Канта имело своеобразное значение. С одной стороны, в предложенном им варианте разрешения антиномии чистого практического разума постулируется существование «нечеловеческих» разумных существ (бога и ангелов). С другой стороны, критерием моральности в его философии выступает разумность, и никакой другой, кроме как присущей людям, пусть и не совсем в «чистом» виде, разумности Кант не предлагает и не допускает. Иными словами, моральными агентами могут быть не только люди, но другие могут быть моральными агентами лишь постольку, поскольку они разумны как люди (вопрос о первенстве этой разумности здесь не затрагивается).

В современных реалиях вопрос о «нечеловеческой» разумности или о разумности «нечеловеческих» моральных агентов получает особый ракурс в силу бурного развития цифровых технологий и появления Систем Искусственного Интеллекта (СИИ), обладающих существенной автономией и особенностями функционирования, которые иногда воспринимаются как признаки моральной агентности. В совокупности эти обстоятельства создают проблемное поле для дискуссий о возможности существования Искусственных Моральных Агентов (ИМА), а также возникающих в связи с этим этических рисков для моральной автономии людей. О важности этого свидетельствует то, что во многих документах этической направленности требования обеспечения и защиты мораль-

ной автономии и достоинства людей рассматриваются как одни из наиболее актуальных [1, 2]. Возникающие в данных контекстах проблемы могут быть объединены в следующие взаимосвязанные группы: (1) возможности и границы в вопросах о моральной автономии ИИ, (2) потенциальное влияние (в том числе и положительное) и риски в отношении моральной автономии людей при взаимодействии с СИИ. При этом в дальнейшем рассмотрении следует учитывать, что рассуждения об автономии не будут затрагивать (а) понимания автономии СИИ в строго функциональном смысле (например, работающий независимо (автономно) робот-пылесос); (б) пусть и широко обсуждаемые, но пока фантастические идеи искусственных как «общего» (сравнимого с человеческим), так и «сверх» (превышающего человеческий) интеллектов.

(1) Исследование этических рисков в связи разработками беспилотных транспортных средств (в частности, автомобилей) породила шутку: «Продавец в автосалоне спрашивает у покупателя о том, с каким моральным алгоритмом должна быть машина: с утилитаризмом Дж. Ст. Милля или этикой долга И. Канта». При всей юмористичности и кажущейся фантастичности подобной ситуации, следует отметить, что она в некотором смысле уже реальна. Так, специалисты в области робототехники уже утверждают, что создание коллаборативных роботов наиболее перспективно, если «моральным» компонентом системы управления является аксиологическое учение с хорошо интерпретируемой системой понятий, например, философская школа Эпикура» [3, с.142]. В упомянутых ранее этических рекомендациях «Этически ориентированный дизайн: подход к обеспечению приоритета благополучия человека с помощью автономных и интеллектуальных систем. Версия 2.0» один из разделов называется «Встраивание ценностей в автономные интеллектуальные системы» [2, с.33–54]. Уже эти обстоятельства свидетельствуют о том, что работы по созданию алгоритмов ИИ с элементами этических концепций уже ведётся.

Оставляя в стороне вопросы о технических и программных возможностях обсуждаемого «встраивания ценностей» в алгоритмы, стоит задуматься о том, насколько такие ИИ могут рассматриваться в качестве

«морально автономных». Одной из пока неразрешимых этических проблем, связанных с созданием ИИ, является его «моральная предвзятость». Существующие алгоритмы воспроизводят или этические предубеждения самих разработчиков, или, если речь идёт о вариантах машинного обучения, существующие стереотипы, прямо или косвенно содержащиеся в данных для обучения. Данные обстоятельства усиливают этические риски дискриминации людей по различным основаниям, что обычно рассматривается как угроза моральной автономии и достоинству людей. Это же характерно и для возможностей машинного самообучения, функционирование которых всё равно не выходит за рамки исходных алгоритмов. Так, играющий в шахматы ИИ сможет сделать нестандартный (изначально не запрограммированный) шахматный ход, но не в состоянии играть в шашки, не говоря уже и о том, чтобы придумать новые правила для любой игры. Современные алгоритмы не способны к тому, что может быть хотя бы метафорично названо самозаконодательством, тем более моральным самозаконодательством. В настоящее время нет оснований полагать, что существующие технологии в какой-то степени могут трактоваться как ИМА. Таким образом, речь может идти только об их функциональной автономии, то есть о способности работать какое-то время без непосредственного вмешательства со стороны людей в соответствии с заранее определенным набором целей, даже если способы их достижения частично формируются в ходе работы ИИ. Говоря словами Канта, алгоритмы ИИ запрограммированы не просто на основе гипотетических императивов, но на их разновидности, которую он называл «техническими императивами» (правилами умения). Иными словами, в таком варианте их даже затруднительно интерпретировать в качестве «прагматических императивов» (советы благоразумия), ориентированных на пользу и благо.

(2) Другим проблемным полем в обсуждаемом контексте является то, как взаимодействие с ИИ может влиять на моральную автономию людей. Анализ существующих в настоящее время дискуссий как среди исследователей, так и в публичном информационном пространстве позволяет выделить следующие наиболее важные проблемные области:

(1) приватность, (2) рекомендательные алгоритмы, (3) моральная «субституция». Их взаимосвязь (при наличии существенных самостоятельных этических аспектов в каждой из этих сфер) может быть представлена следующим образом. Неотъемлемой частью повседневности современных людей стали технологии на основе ИИ, которые отслеживают различные стороны их жизнедеятельности, что предельно сужает сферу частной жизни, которая обычно трактуется как необходимое условие моральной автономии. Даже появилась идея о том, что приватность — это не более чем пережиток, и можно говорить о «смерти приватности» [4]. При этом оставляемые людьми «цифровые следы» рассматриваются в качестве своеобразной платы за предоставление доступа ко многим цифровым услугам. Именно на основе огромного объема персональных и связанных с ними данных происходит обучение ИИ и формируются так называемые рекомендательные алгоритмы, что, вроде бы, освобождает нас от рутинной работы и излишних затрат, помогает принимать более эффективные решения в различных областях и делает нашу жизнь более удобной и комфортной. Моральную озабоченность в этом плане вызывают вопросы полной или частичной замены (субституции) людей интеллектуальными машинами, а также опасности неконтролируемости деятельности алгоритмов, что создает возможности для прямой или косвенной манипуляции, что нарушает свободу и автономию людей. Но ещё большие этические риски связаны с возможностью моральной субституции, то есть перекладывания решения моральных вопросов на алгоритмы ИИ, даже если последние выступают исключительно в виде источника этически значимой информации и только предлагают решения моральных проблем. С одной стороны, напрашивается предположение, что такие технологии могли бы претендовать на цифровой аналог «чистого практического разума», поскольку лишены всех, свойственных людям недостаткам, мешающим принимать морально правильные решения (свобода (автономия) от эмоций и «низшей способности желания», независимость от внешнего влияния и т.д.). С другой стороны, именно в отношении решения моральных проблем современные технологии ИИ демонстрируют свою

максимальную беспомощность. Наглядно это проявляется в публикуемых разработчиками многочисленных дисклеймерах и накладываемых ими алгоритмических ограничениях, в центре внимания которых оказывается отказ от ответственности в отношении большинства этически значимых вопросов. Такое «моральное уклонение» прямо свидетельствует о том, что нравственность по-прежнему является исключительно человеческим делом. Косвенно ставит необходимость, вслед за Кантом, решать вопросы о критериях демаркации нравственности, отделяя ее от внеморальных сфер жизнедеятельности.

Завершая краткое рассмотрение проблемы моральной автономии в контексте развития алгоритмов ИИ, следует отметить, что сформулированный призыв Канта «Имей мужество пользоваться собственным умом!» нисколько не потерял своей актуальности, поскольку отказ от него в пользу технологий с потенциальными элементами ИМА, порождает многочисленные этические риски, в том числе и в отношении самой моральной автономии.

Список литературы

1. Кодекс этики в сфере искусственного интеллекта (Альянс в сфере искусственного интеллекта) [Электронный ресурс] URL: https://ethics.a-ai.ru/assets/ethics_files/2023/05/12/Кодекс_этики_20_10_1.pdf (Дата обращения: 09.01.2024).
2. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2. IEEE, 2017. [Электронный ресурс] URL: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf (Дата обращения: 09.01.2024).
3. Леушина В.В. Об одном подходе к созданию роботов-партнеров как моральных агентов // XV международная конференция «Теоретическая и прикладная этика: Традиции и перспективы – 2023. Разумность. Практичность. Человечность». Санкт-Петербургский Государственный Университет, 16-18 ноября 2023 г. Материалы конференции / Отв.ред. В.Ю. Перов. СПб.: ООО «Сборка», 2023. 240 с., С. 140–142.
4. Froomkin M. The Death of Privacy? // Stanford Law Review, Vol. 52, May 2000, Pp. 1461–1543.

References

1. Kodeks ėtiki v sfere iskusstvennogo intellekta (Al'ians v sfere iskus-stvennogo intellekta) [The Code of Ethics in the Field of Artificial Intelligence (Alliance for Artificial Intelligence)]. Available from: https://ethics.a-ai.ru/assets/ethics_files/2023/05/12/Kodeks_ėtiki_20_10_1.pdf [Accessed 9th January 2024].
2. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2. IEEE, 2017. Available from: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf [Accessed 9th January 2024].
3. Leushina V.V. Ob odnom podkhode k sozdaniiu robotov-partnerov kak moral'nykh agentov [On One Approach to Creating Robot Partners as Moral Agents] // XV International Conference “Theoretical and Applied Ethics: Traditions and Prospects – 2023: Rationality. Practicability. Humanity”. Saint-Petersburg University, Russia. November 16-18, 2023. Conference proceedings / Ed.-in-Chief V.Yu. Perov. SPb.: OOO “Sborka”, 2023. 240 p., Pp. 140–142.
4. Froomkin M. The Death of Privacy? // Stanford Law Review, Vol. 52, May 2000, Pp. 1461–1543.