Человек в информационном обществе

ЭТИЧЕСКОЕ РЕГУЛИРОВАНИЕ РАЗРАБОТКИ И ПРИМЕНЕНИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: ПРОБЛЕМЫ И РЕШЕНИЯ

Петрунин Юрий Юрьевич

Доктор философских наук, профессор
Московский государственный университет имени М.В.Ломоносова, Факультет государственного управления, заведующий кафедрой математических методов и информационных технологий в управлении
Москва, Российская Федерация
petrunin@spa.msu.ru

Кондрашов Павел Евгеньевич

Кандидат технических наук, доцент
Московский государственный университет имени М.В.Ломоносова, Факультет государственного управления, кафедра истории государственного и муниципального управления, ведущий научный сотрудник
Москва, Российская Федерация
kondrashov@spa.msu.ru

Попова Светлана Сергеевна

Кандидат юридических наук, доцент
Московский государственный университет имени М.В.Ломоносова, Факультет государственного управления, кафедра правовых основ управления
Москва, Российская Федерация
popova@spa.msu.ru

Аннотация

В статье рассматриваются вопросы этического регулирования искусственного интеллекта (ИИ). Анализируются проблемы эффективности применения разных этических концепций для ИИ; оптимального выбора раздела этики, соответствующего исследованиям регулирования ИИ; перспективы взаимодействия профессиональной этики и поведенческой экономики; уточнения некоторых ключевых понятий ИИ. Делаются выводы о необходимости преодоления разрыва между академическим сообществом и сообществом практиков-управленцев, разрабатывающих и внедряющих в жизнь механизмы этического регулирования ИИ.

Ключевые слова

искусственный интеллект, этическое регулирование ИИ, прозрачность ИИ, ответственность ИИ, профессиональная этика, поведенческая экономика

Введение

Этическое регулирование ИИ является сегодня важнейшим механизмом минимизации опасностей, связанных со стремительным и неконтролируемым развитием разработок и применением ИИ. Преимуществом этики по сравнению с правовым регулированием является бо́льшая оперативность, гибкость, адаптивность, понятность и доверительность. За последние 10 лет число публикаций в РИНЦ по этике ИИ росло по экспоненте с коэффициентом детерминации R^2=0,98 (вычислено авторами на январь 2025 г.). Отмечая значительный интерес к этическим инструментам регулирования ИИ, следует отметить, что существует определенный разрыв между «официальными» инструментами регулирования [5, 10, 12] и концепциями научного сообщества.

https://doi.org/10.52605/16059921_2025_05_14

В статье рассматриваются некоторые ключевые моменты этого разрыва: выбор теоретических основ этики регулирования ИИ; ключевые понятия этики ИИ; выделение наук, занимающихся вопросами изучения общественной морали; локализация этики для регулирования ИИ.

1 Релевантность этических учений и понятий

В работах по вопросам регулирования ИИ, в основном, используется этика добродетели [10; 6] и деонтология [14; 3]. В национальном стандарте Российской Федерации по ИИ [9] перечислен более широкий круг этических концепций: утилитаризм, деонтология, этика добродетели, но отсутствует, например, этика заботы, которую можно назвать «феминистской этикой», поскольку она возникла под влиянием феминизма [15]. Несмотря на критическое отношение к данному этическому учению в этике заботы есть много интересных и полезных моментов, которые можно использовать в регулировании ИИ на всех стадиях его жизненного цикла.

Среди основных этических требований, предъявляемых к ИИ, на первом месте стоит прозрачность выводов/рекомендаций интеллектуального агента [8, с. 95-97]. Прозрачность ИИ увеличивает доверие к нему. Однако это свойство противоречиво. Во-первых, прозрачность по-настоящему невозможна. Искусственные нейронные сети и машинное обучение не являются строго алгоритмизированными технологиями, результаты их работы невоспроизводимы в принципе. В общем виде прозрачность означает раскрытие контента, описание данных, на которых ИИ обучился, и алгоритмов обработки данных. Во-вторых, прозрачность ИИ, в определенной степени, контекстуальна, имеет свои отраслевые особенности применения: в медицине, в государственном управлении, в образовании. Помимо этого, она может пониматься как обязанность всех пользователей указывать на результаты своей деятельности, при реализации которой был использован ИИ.

Востребованным является также понятие ответственности при применении ИИ. Например, некоторые компании ИИ-сектора уже пытаются избежать ответственности за любое использование их продукции. В размещенных в интернете Условиях OpenAI написано, что «OpenAI не гарантирует надежности, пригодности, качества, соблюдения прав, точности выходных данных, а также ответственности за косвенные, случайные, специальные, последовательные или образцовые убытки» [11]. Вот и первые результаты: 7 января 2025 г. в Лас-Вегасе взорвался автомобиль Tesla Cybertruck. Полиция выяснила, что преступник изготовил взрывное устройство по рецепту ChatGPT. OpenAI парировала: их ИИ-агент поделился информацией, уже имеющейся в открытом доступе в интернете. При этом бот предупредил о незаконности деятельности [21].

В ИИ как технологии выделить конкретное ответственное лицо невозможно. Такое явление в профессиональной этике называется эффектом «ста рук», когда ответственность коллективного действия/деятельности размывается между всеми участниками.

Для решения этой проблемы относительно недавно была разработана концепция «распределенной моральной ответственности» [16; 17]. Её смысл состоит в том, что акцент должен ставиться не на намерениях/мотивациях стейкхолдеров, а на значении возможного воздействия технологии на морально значимые объекты. «Моральная ответственность в рамках распределенного подхода касается всей социотехнической системы в целом, что ведет не к девальвации ответственности отдельных агентов системы, но, напротив, к ее интенсификации, поскольку ответственным в равной мере оказывается каждый агент [4, с.139].

В современном мире этическими проблемами занимается не только философская этика. Существенный вклад в решение этических/моральных проблем внесли экономика и социология. Как известно, отцом и экономики, и этики был великий мыслитель древности Аристотель. Вспомним также, что первая монография Адама Смита, опубликованная еще до знаменитой «Исследование о природе и причинах богатства народов», называлась «Теория нравственных чувств», а сам автор возглавлял одно время кафедру нравственной философии в университете Эдинбурга, выиграв конкурс у самого Дэвида Юма. В XIX в. И. Бентам – тоже крупный экономист – разработал этическое учение утилитаризма, где справедливость и эффективность становятся почти синонимами. В наши дни активное сотрудничество экономики и этики происходит в поведенческой экономике, которая внесла в этику экспериментальный метод, а в экономику – ограничение рационального выбора социальными и моральными стандартами. Здесь, конечно, господствует деонтология И. Канта. Некоторые экономисты даже считают, что равновесие Нэша не более, чем «перевод с языка немецкого идеализма на язык теории игр этического дискурса» [7]. На

январь 2025 г. число публикаций по тематике «этика ИИ» первое место в РИНЦ занимает философия (64 работы), а второе – экономика (49 работ) (рассчитано авторами).

2 Локализация этики для регулирования ИИ

Идентификация проблемного поля этики ИИ чрезвычайно важна при становлении новой научной дисциплины, потому что позволяет не изобретать велосипед, а опираться на уже разработанные методы, модели, терминологию и инструменты разрешения проблем своих предшественников и близких дисциплин. Например, биоэтика не только использует богатый научный багаж медицинской этики, но расширяет и дополняет его. Какой же раздел этики ближе всего к этике ИИ?

В п.4.3.1 Предварительного национального стандарта Российской Федерации «Искусственный интеллект. Обзор этических и общественных аспектов» этика ИИ отнесена к прикладной этике [9]. Такая фокусировка представляется размытой и дискуссионной.

Более разумно сформулирован п. 4.2 указанного выше предварительного стандарта, где «приведен набор заинтересованных сторон, участвующих в разработке и использовании системы ИИ, а также описано участие различных заинтересованных сторон ИИ в цепочке создания стоимости системы ИИ, которые включают поставщика ИИ, производителя ИИ, заказчика ИИ, партнера ИИ и субъекта ИИ, в том числе различные второстепенные роли этих заинтересованных сторон» [9]. Ближе всего данной проблематикой занимается т.н. профессиональная этика, начало которой положено работой социолога Э. Дюркгейма [13]. Например, проблема передачи ответственности юридического лица, подписывающего какие-либо правила и кодексы этики в сфере ИИ, сотрудникам этой организации, уже анализировалась в профессиональной этике. «Профессиональные этические кодексы имеют дело с индивидами и индивидуальным поведением. Это отличает англо-американскую и повторяющую ее континентальную модели от старых немецкой и французской моделей. До 1789 года французские профессии строили свои этические кодексы на основе корпоративных обязательств, возложенных на них государством. В условиях индивидуализма из этических кодексов исчезают положения, описывающие обязательства профессии в целом» [12, с. 860-861]. Подписанты «Декларации об ответственной разработке и использовании сервисов на основе генеративного искусственного интеллекта» [5] – крупнейшие российские компании и ведущие научные и образовательные организации нашей страны – не предусмотрели, как передать эту ответственность своим сотрудникам.

Ученые из ВШЭ отмечают, что «ИИ-системы, как и любые другие технические системы, не имеют внутренне присущих им свойств в области морали – эти свойства характерны для процессов жизненного цикла систем. Можно судить об этичности процедуры сбора данных при создании ИИ-системы, этичности применения систем для решения той или иной прикладной задачи, этичности интерпретации и использования результатов обработки данных системой и т.п.» [10, с. 147]. Фактически, каждый этап – это отдельная профессия, в том числе со своей профессиональной этикой и моралью. Соответственно, и этика ИИ включает в себя несколько локальных профессиональных этик, отличающихся от других своими специфическими проблемами, терминологией, инструментами и методами, но использующими общие этические теории. При этом могут возникать «и сложности взаимодействия этосов разных профессий, и вопросы использования механизмов этической регуляции в целях недобросовестной конкуренции» [2, с. 82].

Заключение

Как видно из анализа существующих проблем, эффективность этического регулирования ИИ ослабляется из-за разрыва академического сообщества и сообщества практиков-управленцев, разрабатывающих и внедряющих в жизнь «действующие» механизмы регулирования (этические кодексы, декларации и т.п.). Надо отметить, что аналогичные коллизии существуют и зарубежом. «Наш обзор показывает, что политические и экономические последствия деловой практики ИИ в значительной степени недопредставлены в руководящих принципах этики ИИ»… «системы ИИ, по-прежнему серьезно подорваны конкурентными и спекулятивными нормами и другими вредными деловыми практиками» [19, с. 389]; «Систематически рассматривая исследовательские работы, упоминающие этические термины в рамках и инструментах xAI… мы наблюдаем ограниченное и часто поверхностное взаимодействие с этическими теориями» [20, с. 26]; «во многих существующих академических исследованиях отсутствует практический и объективный вклад в разработку этических систем ИИ» [21]. Взаимные упреки теоретиков и практиков должны

стимулировать разные стороны к объединению усилий, цель которых – позитивное использования ИИ для блага человека, государства и общества.

Литература

1. Алексеев А.П., Алексеева И. Ю. Статус этических кодексов в этике искусственного интеллекта // Информационное общество. 2024. № 4. С. 43-49. DOI: 10.52605/16059921_2024_04_43
2. Алексеева И. Ю. Этика искусственного интеллекта как прикладная этика // Философия и общество. 2024. No 3. С. 69–85. DOI: 10.30884/jfio/ 2024.03.06.
3. Антипов А. В. Искусственные моральные агенты: деонтология и моральный тест Тьюринга // Koinon. 2024. Т. 4. № 1–2. С. 9–17. DOI: 10.15826/koinon.2024.04.1.2.001
4. Глуховский А.С., Дурнев А.Д., Чирва Д.В. Распределенная моральная ответственность в сфере искусственного интеллекта // Этическая мысль. 2024. Т. 24. № 1. С. 129–143. DOI: 10.21146/2074-4870-2024-24-1-129-143.
5. Декларация об ответственном генеративном ИИ. URL: https://ethics.a-8ai.ru/genai-declaration.
6. Кудряшова В.К. Может ли искусственный интеллект быть «этичным»? // Этическая мысль. 2024. Т. 24. № 1. 101–114. DOI: 10.21146/2074-4870-2024-24-1-101-114.
7. minski_gaon. Нужна ли сейчас этика вообще? И есть ли она? URL: https://methodology-ru.livejournal.com/197871.html? Aug. 22nd, 2013
8. Петрунин Ю.Ю. Развитие концепции социального искусственного интеллекта // Вестник Московского университета. Серия 21. Управление (государство и общество). 2023. № 1. С. 93-112.
9. ПНСТ 840-2023. Искусственный интеллект. Обзор этических и общественных аспектов. Федеральное агентство по техническому регулированию и метрологии. М., Российский институт стандартизации. 2023.
10. Углева А.В., Шилова В.А., Карпова Е.А. Индекс «этичности» систем искусственного интеллекта в медицине: от теории к практике // Этическая мысль. 2024. Т. 24. № 1. С. 144–159. DOI: 10.21146/2074-4870-2024-24-1-144-159
11. Условия использования Open AI. URL: https://openai.com/terms/
12. Abbot A. Professional Ethics // The American Journal of Sociology. 1983. № 5. Pp. 855 – 885.
13. Durkheim E. Professional Ethics and Civic Morals. Glencoe. Il.: Free Press. 1958. 228 p.
14. Mougan C., Brand J. Kantian Deontology Meets AI Alignment: Towards Morally Grounded Fairness Metrics. arXiv:2311.05227. 2023.
15. Gilligan C. In a Different Voice: Psychological Theory and Women's Development. Cambridge, Mass.: Harvard University Press. 1982. Pp. 24-39.
16. Floridi L. Distributed Morality in an Information Society // Science and Engineering Ethics. 2012. Vol. 19. Pp. 727–743.
17. Floridi L. Faultless Responsibility: on the Nature and Allocation of Moral Responsibility for Distributed Moral Actions // Philosophical Transactions of the Royal Society. Series A, 2016. № 374 (2083). Pp. 1–13.
18. Attard-Frost B., De los Ríos A., Walters D.R. The ethics of AI business practices: a review of 47 AI ethics guidelines // AI Ethics. 2023. No 3. Pp. 389–406. https://doi.org/10.1007/s43681-022-00156-6
19. Nannini L., Manerba M. M., Beretta I. Mapping the landscape of ethical considerations in explainable AI research // Ethics and Information Technology. 2024. Pp. 26:44. https://doi.org/10.1007/s10676-024-09773-7
20. Palumbo G., Carneiro D., Alves V. Objective metrics for ethical AI: a systematic literature review // International Journal of Data Science and Analytics https://doi.org/10.1007/s41060-024-00541-w. Accepted: March 2024
21. Tucker E., Green Beret who exploded Cybertruck in Las Vegas used AI to plan blast // CNN. January 7, 2025. URL: https://edition.cnn.com/2025/01/07/us/las-vegas-cybertruck-explosion-livelsberger/index.html

ETHICAL REGULATION OF THE DEVELOPMENT AND APPLICATION OF ARTIFICIAL INTELLIGENCE: PROBLEMS AND SOLUTIONS

Petrunin Yurij Yurievich

DSc in philosophy, professor
Lomonosov Moscow State University, School of public administration, Department of mathematical methods and information technology in management, chairman
Moscow, Russian Federation
petrunin@spa.msu.ru

Kondrashov Pavel Evgenievich

Candidate of technical sciences, associate professor
Lomonosov Moscow State University, School of public administration, Department of history of state and municipal administration, lead researcher
Moscow, Russian Federation
kondrashov@spa.msu.ru

Popova Svetlana Sergeevna

PhD in law, associate professor
Lomonosov Moscow State University, School of public administration, Department of legal foundations for public administration
Moscow, Russian Federation
popova@spa.msu.ru

Abstract

The article considers the issues of ethical regulation of artificial intelligence (AI). The article analyzes the problems of the effectiveness of applying different ethical concepts to AI; the optimal choice of a section of ethics corresponding to studies of AI regulation; prospects for the interaction of professional ethics and behavioral economics; clarification of key concepts of AI. In conclusion, the article draws conclusions about the need to overcome the gap between the academic community and the community of practicing managers who develop and implement mechanisms for the ethical regulation of AI.

Keywords

artificial intelligence, ethical regulation of AI, AI transparency, AI responsibility, professional ethics, behavioral economics

References

1. Alekseev A. P., Alekseeva I. Yu. Status e`ticheskix kodeksov v e`tike iskusstvennogo intellekta // Informacionnoe obshhestvo. 2024. No 4. S. 43–9. Izvlecheno ot http://infosoc.iis.ru/article/view/702.
2. Alekseyeva I. Yu. E`tika iskusstvennogo intellekta kak prikladnaya e`tika // Filosofiya i obshchestvo = Philosophy and Society. 2024. No. 3. Pp. 69–85. DOI: 10. 30884/jfio/2024.03.06.
3. Antipov A. V. Iskusstvenny`e moral`ny`e agenty`: deontologiya i moral`ny`j test T`yuringa // Koinon. 2024. T. 4. No 1–2. C. 9–17. DOI: 10.15826/koinon.2024.04.1.2.001
4. Glukhovskii A. S., Durnev A. D., Chirva D. V. Raspredelennaya moral`naya otvetstvennost` v sfere iskusstvennogo intellekta // E`ticheskaya my`sl`. 2024. Vol. 24, No. 1, pp. 129-143. DOI: 10.21146/2074-4870-2024-24-1-129-143
5. Deklaraciya ob otvetstvennom generativnom AI. M. 2024.
6. Kudriashova V. K. Mozhet li iskusstvenny`j intellekt by`t` «e`tichny`m»? // E`ticheskaya my`sl`. 2024. Vol. 24, No. 1, pp. 101–114. DOI: 10.21146/2074-4870-2024-24-1-101-114
7. minski_gaon. Nuzhna li sejchas e`tika voobshhe? I est` li ona?? URL: https://methodology-ru.livejournal.com/197871.html? Aug. 22nd, 2013
8. Petrunin Yu.Yu. Razvitie koncepcii social`nogo iskusstvennogo intellekta // Vestnik Moskovskogo universiteta. Seriya 21. Upravlenie (gosudarstvo i obshhestvo). 2023. Vol. 20. № 1. P. 93-112.

9. PNST 840-2023. Iskusstvenny`j intellekt. Obzor e`ticheskix i obshhestvenny`x aspektov. Federal`noe agentstvo po texnicheskomu regulirovaniyu i metrologii. M., Rossijskij institut standartizacii. 2023.
10. Ugleva A. V., Shilova V. A., Karpova E. A. Indeks «e`tichnosti» sistem iskusstvennogo intellekta v medicine: ot teorii k praktike // E`ticheskaya my`sl`. 2024. Vol. 24, No. 1, pp. 144–159. DOI: 10.21146/2074-4870-2024-24-1-144-159.
11. Conditions of use. Open AI. URL: https://openai.com/terms/
12. Abbot A. Professional Ethics // The American Journal of Sociology. 1983. №. 5, p. 855 – 885.
13. Durkheim, E. Professional Ethics and Civic Morals. Glencoe, Ill.: Free Press. 1958. 228 p.
14. Mougan C., Brand J. Kantian Deontology Meets AI Alignment: Towards Morally Grounded Fairness Metrics. arXiv:2311.05227. 2024.
15. Gilligan C. In a Different Voice: Psychological Theory and Women's Development. Cambridge, Mass.: Harvard University Press. 1982. Pp. 24-39.
16. Floridi L. Distributed Morality in an Information Society. – Science and Engineering Ethics. 2012. Vol. 19, pp. 727–743.
17. Floridi L. Faultless Responsibility: on the Nature and Allocation of Moral Responsibility for Distributed Moral Actions // Philosophical Transactions of the Royal Society, Series A, 2016. № 374 (2083), pp. 1–13.
18. Attard-Frost B., De los Ríos A., Walters D.R. The ethics of AI business practices: a review of 47 AI ethics guidelines. // AI Ethics. 2023. № 3, 389–406. https://doi.org/10.1007/s43681-022-00156-6
19. Nannini L., Manerba M. M., Beretta I. Mapping the landscape of ethical considerations in explainable AI research // Ethics and Information Technology. 2024. 26:44. https://doi.org/10.1007/s10676-024-09773-7.
20. Palumbo G., Carneiro D., Alves V. Objective metrics for ethical AI: a systematic literature review // International Journal of Data Science and Analytics https://doi.org/10.1007/s41060-024-00541-w. Accepted: March 2024.
21. Tucker E., Green Beret who exploded Cybertruck in Las Vegas used AI to plan blast // CNN. January 7, 2025. URL: https://edition.cnn.com/2025/01/07/us/las-vegas-cybertruck-explosion-livelsberger/index.html