DOI: 10.22394/3034-2813-2024-4-56-67 EDN: YLOUUO

Дата поступления статьи: 09.09.2024
Дата поступления рецензии: 24.09.2024
Дата принятия статьи к публикации: 01.11.2024

Искусственный интеллект и социальное неравенство: философско-правовой аспект[1]

Рыбин А. И.[1, *], **Чащухин Е. О.**[1, **]
[1] Национальный исследовательский университет «Высшая школа экономики» (Москва, Российская Федерация)
[*] e-mail: airybin@hse.ru
[**] e-mail: eochaschukhin@hse.ru

Аннотация
В настоящей статье ставится цель проанализировать влияние искусственного интеллекта на социальное неравенство. Показано, что искусственный интеллект не только порождает блага, но в том числе является инструментом нарушения прав человека, увеличивающим социальное расслоение как на уровне отдельных обществ, так и в мировом масштабе. При помощи формально-юридического и сравнительно-правового методов показано, что существующая в настоящий момент практика применения искусственного интеллекта в различных сферах права не всегда достигает заложенных авторами целей, а иногда становится причиной дискриминации и воспроизводства социального неравенства. В работе продемонстрировано, что социальное неравенство и другие проблемы, связанные с внедрением искусственного интеллекта, активно осмысляются в рамках этики искусственного интеллекта. Проанализированы концепции этики искусственного интеллекта, а также ряд этических кодексов, которые приняты в ряде технологических компаний, занимающихся искусственным интеллектом. Показано, что основной причиной появления социального неравенства в этой сфере является не отсутствие этических принципов программ, действующих на основе искусственного интеллекта, а недостаточная формализация самих этических принципов на машинный язык. Показано, что правовое регулирование данной сферы, как правило, носит рекомендательный характер и исходит от корпораций, а не от государств.
Ключевые слова: искусственный интеллект, цифровое неравенство, социальная стратификация, этика, марксистская критика неравенства.
Для цитирования: Рыбин А. И., Чащухин Е. О. Искусственный интеллект и социальное неравенство: философско-правовой аспект // Теоретическая и прикладная юриспруденция. 2024. № 4 (22). С. 56–67. DOI: 10.22394/3034-2813-2024-4-56-67. EDN: YLOUUO

Artificial Intelligence and Inequality: A Legal Aspect[2]

Rybin A. I.[1,*], **Chashhukhin E. O.**[1,**]
[1] National Research University Higher School of Economics (Moscow, Russian Federation)
[*] e-mail: airybin@hse.ru
[**] e-mail: eochaschukhin@hse.ru

[1] В данной научной работе использованы результаты проекта «Правовые механизмы преодоления неравенства», выполненного в рамках Программы фундаментальных исследований НИУ ВШЭ в 2024 г.
[2] The results of the project “Legal mechanisms of overcoming inequality”, carried out within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE University) in 2024, are presented in this work.

Abstract

The purpose of this article is to analyze the impact of artificial intelligence (hereinafter also referred to as AI) on social inequality. It is evident that AI not only brings about advantages but also serves as a means of infringing upon human rights, exacerbating social stratification at both the level of individual societies and on a global scale. Through the application of formal legal and comparative legal methodologies, it becomes apparent that the current practices of utilizing AI in various legal domains often fall short of achieving the intended objectives, sometimes even serving as a catalyst for discrimination and the perpetuation of social inequality. The paper underscores the active engagement of scholars and practitioners in addressing social inequality and other challenges associated with AI through the lens of AI ethics. The concept of AI ethics is explored, along with a critical analysis of ethical frameworks adopted by several technology companies operating in the field of artificial intelligence. It has been demonstrated that the primary cause of the emergence of social disparity in the realm of AI is not a dearth of ethical tenets in AI-driven algorithms, but rather a lack of adequate formalization of these ethical principles themselves into machine-readable language. It has been shown that the legal regulations in this domain, typically, are advisory in nature and emanate from corporations rather than from state institutions.

Keywords: artificial intelligence, digital inequality, social stratification, ethics, Marxist critique of inequality.

For citation: Rybin, A. I., Chashhukhin, E. O. Artificial Intelligence and Inequality: A Legal Aspect. *Theoretical and Applied Law.* No. 4 (22). Pp. 56–67. (In Russ.) DOI: 10.22394/3034-2813-2024-4-56-67

Введение

Искусственный интеллект (далее — ИИ) является одним из самых обсуждаемых и влиятельных новшеств как в технической, так и в гуманитарной сфере. Это закономерно, поскольку те изменения, причиной которых становится ИИ, качественно изменяют жизнь людей, а также влияют на продвижение отдельных отраслей экономики, особенно технологических. Вместе с тем внедрение ИИ в повседневные практики не всегда происходит по заранее продуманному сценарию.

Даже если функционирование некоторых юнитов ИИ и возможно предугадать с технической стороны, изменения в обществе, происходящие из-за внедрения ИИ в повседневный обиход, зачастую непредсказуемы. Непредсказуемость является отличительной чертой ИИ, поскольку его функционирование происходит по принципу «черного ящика», в котором мы можем видеть лишь состояние данных на «входе» и результат их переработки на «выходе». Более того, учитывая тот факт, что данные, на которых обучается ИИ, в ряде случаев могут оказаться некорректными, последствием использования такого юнита ИИ может стать нарушение прав человека. Одним из недостаточно проанализированных, по нашему мнению, рисков использования ИИ является увеличение социального расслоения в обществе.

Так, существует распространенное мнение, что ИИ, в силу своей независимости от различных человеческих пороков, может стать причиной сокращения неравенства. Однако в том случае, если юнит ИИ обучался на некорректной выборке данных, он неминуемо будет воспроизводить негативные паттерны, способствующие увеличению неравенства. В мировой практике внедрения ИИ уже существует ряд кейсов, подтверждающих этот тезис. Они будут подробно рассмотрены ниже.

Подобная непредсказуемость вместе со значительными вычислительными способностями, превосходящими способности человека, обнажает проблематику определения этических принципов функционирования ИИ, которые призваны обезопасить взаимодействие ИИ с человеком.

Обзор существующих концепций в сфере этики ИИ

Тематика исследований в сфере этики искусственного интеллекта в целом делится на два больших блока. Первый, более крупный, связан с вопросами поведения разработчиков, производителей

и операторов машин, функционирующих на основе ИИ, и нацелен на то, чтобы свести к минимуму тот ущерб, который могут нанести обществу ИИ из-за ненадлежащего программирования, неподходящего применения или неправильного использования из-за игнорирования этических аспектов. В рамках этого же блока рассматривается этика применения роботов и ИИ и называется, как правило, «этикой искусственного интеллекта» или «этикой роботов» (AI ethics, robot ethics)[3]. В этом же блоке разрабатываются этические принципы, стандарты и нормативные акты, призванные установить консенсус относительно этических правил разработки ИИ[4].

И если первый блок исследований этики ИИ связан с регулированием действий человека, то второй — с вопросами о том, какое поведение роботов с ИИ следует считать этичным и как они должны регулировать свое поведение. Эта область именуется «машинная этика» (machine ethics), она подразумевает использование как философских подходов, так и методов технических наук[5]. Философское рассмотрение актуализируется для решения фундаментальных вопросов: может ли машина (робот, ИИ) действовать этично? Если может, то как этика должна определять ее поведение? Какие этические концепции релевантны для машинной адаптации?

К философским вопросам, наряду с более общими, относится следующий: должно ли общество возлагать моральную ответственность на роботов или машины, действующие на основе ИИ? Этот вопрос граничит и с юридической сферой — в части возможности наложения юридической ответственности на юнитов ИИ[6]. К техническим относятся вопросы о том, как, например, запрограммировать этическую машину. Эти сферы в области этики ИИ взаимосвязаны. Инженеры, занимающиеся разработкой «этичных» машин, нуждаются в советах философов при определении соответствующих этических правил и ценностей, закладываемых при программировании ИИ.

В целом идея о том, что роботы должны воздерживаться от причинения вреда людям и предотвращать нанесение вреда человеком человеку, то есть вести себя «этично», имеет длительную историю, причем не только в научной литературе, но и в научной фантастике. Еще в 1940-х гг., в период создания первых ЭВМ, в рассказе «Хоровод» Айзек Азимов сформулировал ныне хорошо известные законы робототехники[7]:

1) Робот не может причинить вред человеку или своим бездействием допустить, чтобы человеку был причинен вред.

2) Робот должен повиноваться всем приказам, которые дает человек, кроме тех случаев, когда эти приказы противоречат Первому Закону.

3) Робот должен заботиться о своей безопасности в той мере, в которой это не противоречит Первому или Второму Законам.

Хотя перечисленные «законы» в изложении А. Азимова были не более чем художественным вымыслом, пускай и логически связанным, они были не единственным предложением по этическому ограничению ИИ в сфере машинной этики. В последующие годы предпринимались (и в настоящее время продолжают предприниматься) попытки этического регулирования действий ИИ. В книге «Моральные машины — обучая роботов отличать правильное от неправильного» У. Уоллах и К. Аллен изложили философские основы машинной этики, введя термин «Искусственный моральный агент» (Artificial Moral Agent, AMA)[8].

Помимо проблемы причинения вреда жизни или здоровью человека вследствие недостаточной продуманности этических принципов взаимодействия ИИ с человеком, возникают и менее глобальные,

---

[3] *Winfield A., Michael K., Pitt J., Evers V.* (2019) Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems. *Proceedings of the IEEE*, no. 107 (3), p. 510. DOI: 10.1109/JPROC.2019.2900622

[4] *Boden M. [et al.]* (2017) Principles of robotics: Regulating robots in the real world. *Connection Sci*, vol. 29 (2), p. 124. DOI: https://doi.org/10.1080/09540091.2016.1271400

[5] *Winfield A. [et al.]* Ibid. P. 510.

[6] См. например: *Апостолова Н. Н.* Ответственность за вред, причиненный искусственным интеллектом // Северо-Кавказский юридический вестник. 2021. № 1. С. 112–119. DOI: 10.22394/2074-7306-2021-1-1-112-119

[7] *Murphy R., Woods D.* (2009) Beyond Asimov: The Three Laws of Responsible Robotics. Intelligent Systems, IEEE, no. 24 (4), p. 14. DOI: 10.1109/MIS.2009.69

[8] См.: *Wallach W., Allen C.* (2009) Moral Machines: Teaching Robots Right from Wrong. *Oxford Univ. press*. 288 p.

но не менее важные вопросы. В частности, вопросы функционирования социальных институтов, изменений в расстановке социальных статусов и взаимодействия социальных ролей. В настоящее время уже известен ряд проблемных сюжетов, которые являются с описанной точки зрения наиболее рискованными: вопросы защиты человеческого достоинства (например, в части использования технологии deepfake), охраны персональных данных, неприкосновенности частной жизни, дискриминации и иных проявлений неравенства (digital divide). Представляется, что эти проблемы можно классифицировать в зависимости от того, возможно ли их устранить с помощью только технических или только социальных решений. Например, причины проблем утечки данных и неприкосновенности частной жизни находятся в большей степени в технической плоскости. Проблемы неравенства, возникающие из-за внедрения ИИ, в большей мере находятся в социальной сфере, поэтому исследователям в области гуманитарных и социальных наук следует, в первую очередь, рефлексировать именно их.

Теоретическое объяснение неравенства, порождаемого внедрением ИИ

Прежде чем мы перейдем к описанию реальных кейсов, приводящих в результате действий алгоритмов ИИ к социальному неравенству, имеет смысл обратиться к истории вопроса о причинах такого неравенства. Релевантную позицию по аналогичному вопросу высказывал в свое время К. Маркс в связи с внедрением в жизнь технических изобретений во времена первой промышленной революции.

Отправной точкой для Маркса стали слова Дж. Милля: «Сомнительно, чтобы все сделанные до сих пор механические изобретения облегчили труд хотя бы одного человеческого существа». Маркс продолжил его мысль, скорректировав ее. Он подчеркнул, что применяемые капиталистами машины изначально не предназначались для такой цели. Их цель состоит в том, чтобы «удешевлять товары, сокращать ту часть рабочего дня, которую рабочий употребляет на самого себя, и таким образом удлинять другую часть его рабочего дня, которую он даром отдает капиталисту»[9]. Таким образом, машины, внедряемые в ходе промышленной революции, оказываются не средством, которое как-либо улучшает положение эксплуатируемого класса, но, в первую очередь, становятся средством производства прибавочной стоимости. Дополнительно следует отметить, что внедрение технических новаций увеличивает проявляющееся экономическое неравенство и по другим причинам: перенесение стоимости этих новаций на товары и, как следствие, их удорожание, а потому меньшая доступность для конечного потребителя; интенсификация труда; использование детского труда; ненормированный рабочий график и т. п.

Несмотря на то что в настоящее время культурный, исторический и социально-экономический контексты существенно изменились, тем не менее аргументы, выдвигаемые «за» и «против» внедрения ИИ в сферу производства, (например, о сокращении затрачиваемого на работу времени) представляются аналогичными тому, о чем пишет К. Маркс в процитированном отрывке. Социальное расслоение не только не преодолевается внедрением машин в процесс производства, но наоборот усиливается. С одной стороны, нейросети, которые могут генерировать тексты, видео, изображения, звуки, анализировать большие данные, действительно сокращают рабочее время тех людей, которые этим занимаются; с другой — эти же нейросети ставят под угрозу существование творческих профессий, превращая их в достаточно механистический промпт-инжиниринг, сводимый к заданию команд для ИИ.

В дополнение к экономическим причинам неравенства, порождаемого ИИ, исследователи довольно часто отсылают и к культурным. Регулярно в контексте внедрения ИИ обнажается «проблема отчуждения»[10]. Обобщенное понимание отчуждения по Марксу состоит в превращении результатов человеческой деятельности в самостоятельную сущность, довлеющую над ним, враждебную ему, превращающую его из человека в объект общественного процесса[11]. Типичным примером отчуждения яв-

---

[9] *Маркс К.* Капитал. Т. 1. М.: Государственное издательство политической литературы, 1952. 794 с.

[10] См.: *Юницкий А. Э., Петров Е. О.* Искусственный интеллект и отчуждение человека от разума: причины, механизмы, последствия // Сборник материалов V международной научно-технической конференции «Безракетная индустриализация ближнего космоса: проблемы, идеи, проекты», № 1, 2022, С. 138–143. EDN: MCIVZU; *Sidorkin A. M.* (2024) Embracing liberatory alienation: AI will end us, but not in the way you may think. AI & Soc (2024). DOI: https://doi.org/10.1007/s00146-024-02019-6

[11] *Маркс К.* Экономическо-философские рукописи 1844 г. // К. Маркс, Ф. Энгельс. Собр. соч. Т. 42. М.: Политиздат, 1974. С. 94.

ляются вынесенные ИИ суждения о действиях того или иного человека, который «в глазах» ИИ является не человеком, а некой сущностью, набором параметров, отчужденной тенью человека. Более подробно такие случаи будут рассмотрены ниже.

Еще одной проблемой, обсуждаемой в контексте взаимодействия с ИИ, является «проблема культурной гегемонии». Концепция культурной гегемонии берет свое начало в работах итальянского неомарксиста Антонио Грамши. Согласно Грамши, гегемония — это политическая сила (power), которая базируется не только на власти (authority) и вооруженных силах (военных, полиции и т. д.), но и на общественном мнении, формируемом моральными и интеллектуальными авторитетами[12]. В контексте рассматриваемой темы данная проблема обнаруживается во взаимосвязи с выборками данных, которые анализирует ИИ. Над этими выборками довольно часто довлеет культурная модель-гегемон, со своими достоинствами и недостатками, которая препятствует объективному анализу данных, а потому довольно часто результаты такого анализа воспроизводят существующие дискриминационные предрассудки.

Если говорить в целом, марксистская логика неравенства эксплуатируемых и эксплуататоров может применяться в контексте цифрового неравенства (digital divide) как на индивидуальном, так и на глобальном уровне. Так, сам факт разной доступности к технологиям ИИ, которая коррелирует с возрастом, уровнем навыков и возможностью обладать техническими средствами, предопределяет различие в степени интегрированности в глобальное цифровое пространство. Кроме того, те, кто обладают средствами ИИ (теми, которые не находятся в общем доступе), имеют существенно лучшие условия для производства цифровых работ и услуг. Во всяком случае они могут затрачивать значительно меньше временных и человеческих ресурсов[13].

Итак, мы кратко очертили теоретические предпосылки, которыми обуславливается неравенство при введении ИИ в повседневный обиход. Далее более подробно рассмотрим причины, в соответствии с которыми ИИ становится причиной социального неравенства.

Причины, по которым ИИ порождает социальное неравенство

Искусственный интеллект может способствовать дискриминации, что становится всё более актуальной темой в литературе по социальным наукам. Обратимся к нескольким факторам, играющим ключевую роль в этом процессе.

Основная причина — некорректность данных, тесно связанная с проблемой культурной гегемонии, на основе которых обучается ИИ. Это большие объемы данных, которые могут содержать дискриминационные паттерны. Эти данные могут быть собраны из различных источников, включая социальные медиа, профили пользователей в социальных сетях, статистические базы данных и др., выборка которых по критериям этичности часто не производится. Поэтому если в таких данных присутствуют системные стереотипы, ИИ автоматически воспримет их и будет воспроизводить.

Например, если алгоритм анализирует данные о найме на работу и в них преобладает информация о мужчинах, он может неправомерно недооценивать женщин. Похожие случаи уже имели место на практике. Известно, что в ряде технологических компаний, которые в последнее десятилетие для анализа входящих резюме внедряли ИИ, оказывалось, что любые достижения женщин-кандидатов, вроде успехов в женском спорте или наличия женских хобби, маркировались ИИ как недостатки, а не как достоинства[14].

Так получалось по нескольким причинам: во-первых, в технологических компаниях традиционно преобладает мужской контингент; во-вторых, именно на таких данных обучался алгоритм ИИ. И поскольку данные об успехах в женском спорте или о женских хобби полностью или почти отсутствовали в проанализированных анкетах, полученная информация была распознана как характеристика неуспешных

---

12 *Gramsci A.* (1974) Selections from the prison notebooks of Antonio Gramsci. *International Publishers, Political Science*. 483 p.

13 *Bentley S. V., Naughtin C. K., McGrath M. J. [et al.]* (2024) The digital divide in action: how experiences of digital technology shape future relationships with artificial intelligence. AI and Ethics, no. 3, pp. 7–8. DOI: https://doi.org/10.1007/s43681-024-00452-3

14 *Xinyu C.* (2023) Gender Bias in Hiring: An Analysis of the Impact of Amazon's Recruiting Algorithm. *Advances in Economics, Management and Political Sciences*, no. 23, p. 136. DOI: https://doi.org/10.54254/2754-1169/23/20230367

кандидатов. Таким образом, при использовании информации, основанной, например, на предрассудках, как в случае с технологическими компаниями, ИИ может неосознанно поддерживать существующие стереотипы, что, в свою очередь, может оказывать долгосрочное воздействие на общественное мнение и отношения между различными социальными группами.

Осмысление этого кейса показало, что в данном случае работает принцип «мусор на входе — мусор на выходе»[15]. Этот принцип в некотором смысле соотносится с принципом функционирования ИИ как «черного ящика». Несмотря на то, что неизвестно, как именно ИИ обрабатывает анализируемые данные, можно практически наверняка сказать, что, если предоставить ИИ нерепрезентативную выборку данных, то он, основываясь лишь на этих данных, сможет выдать только некорректный результат[16].

Однако сферой трудоустройства случаи неравенства, создаваемого или воспроизводимого ИИ, не исчерпываются. Другой пример негативного воздействия касается аспектов правоохранительной деятельности. В частности, ИИ может быть весьма полезным в деле профилактики и предотвращения совершения преступлений. Например, в настоящее время существует система «умных городов», в которую интегрированы технологии распознавания лиц, автоматической фиксации правонарушений на транспорте и прочих средств фиксации. ИИ помогает проанализировать данные о наиболее неблагоприятных районах и выявить лиц, склонных к совершению правонарушений. Это помогает координировать действия правоохранительных органов и служит более тонкому ранжированию текущих задач сотрудников правоохранительных органов[17]. Однако этот процесс не застрахован от возникновения «когнитивных ошибок». Данные о совершённых преступлениях релевантны только в случае существования аналогичных обстоятельств. Однако в случае если обстоятельства меняются, ИИ будут продолжать воспроизводить выявленные ранее закономерности. Так, люди, живущие в районах с более низким доходом или уровнем образования, по умолчанию станут объектом внимания полиции. Эта ситуация может привести к так называемому самосбывающемуся пророчеству. Например, обучение ИИ на основе данных о задержании представителей определенных рас или национальностей может привести к формированию модели, ошибочно интерпретирующей это как свидетельство большей предрасположенности к преступности таких категорий граждан, что, в свою очередь, становится причиной воспроизводства дискриминационных стереотипов[18].

Кроме того, существует ряд препятствий, которые мешают эффективности функционирования ИИ: низкая точность прогнозирования, ограниченный круг преступлений, которые можно предсказать, высокая стоимость оборудования и программного обеспечения, некорректный ввод данных и высокий риск предвзятого характера некоторых программ[19].

Другой сферой потенциального применения ИИ является помощь в принятии решений, касающихся реабилитации правонарушителей. Например, в контексте принятия решений об условно-досрочном освобождении от отбывания наказания в виде лишения свободы или в иных вопросах «исправления» лиц, совершивших преступления, связанных с ограничением их прав. Так, ИИ мог бы помочь в анализе массива данных, связанных с поведением преступников в условиях отбывания наказания и допущенных после этого рецидивов, и это могло бы в дальнейшем послужить основанием для более точного вывода о возможном последующем поведении лица, совершившего правонарушение.

Однако и в этом свете ИИ может становиться причиной воспроизводства социального неравенства. Известен случай, когда Гленну Родригесу, в молодости совершившему вооруженное ограбление автосалона, в котором погиб один человек, спустя 25 лет отказали в досрочном освобождении, несмотря на законные основания для такого освобождения и высокие показатели по программе реабилитации.

---

15 *Шталь Б. К. [и др.]* Этика искусственного интеллекта: кейсы и варианты решения этических проблем / Б. К. Шталь, Д. Шредер, Р. Родригес; пер. с англ. И. Кушнаревой; под науч. ред. А. Павлова; Нац. исслед. ун-т «Высшая школа экономики». М.: Изд. дом Высшей школы экономики, 2024. 200 с. С. 26.

16 *Mahadi N., Bakari A. H. A., Baskaran S., Maideen M. H.* (2022) How The Pandemic Has Disrupted and Changed Hiring // International Journal of Academic Research in Business and Social Sciences. No. 12 (10). P. 1335.

17 *Шталь Б. К.* Указ соч. С. 30–31.

18 *Hung T. W., Yen C. P.* (2023) Predictive policing and algorithmic fairness. *Synthese*. Pp. 205–206.

19 *Mugari I., Obioha E.* (2021) Predictive Policing and Crime Control in the United States of America and Europe: Trends in a Decade of Research and the Future of Predictive Policing. *Social Sciences*, no. 10 (234), p. 1. DOI: 10.3390/socsci10060234

Причиной, по которой произошел отказ, стала программа COMPAS, которая, по всей видимости, также могла учитывать мнения сотрудников мест исполнения наказания, проявивших в отношении данного осужденного расовые предрассудки[20].

Более подробно исследовавшие вопрос применения таких технологий в пенитенциарной системе Дж. Вилласенор и В. Фогго приводят также несколько иных инцидентов, связанных с использованием этой программы и вызвавших общественный резонанс в США[21]. В связи с приведенными примерами следует отметить, что методология, используемая для сбора данных, также может стать причиной предвзятости.

Аналогичный механизм мог бы действовать и на уровне уголовного процесса, когда необходимо принимать решение об избрании меры пресечения. В уголовном процессе меры пресечения ранжируются по строгости и степени ограничения прав обвиняемого, характеристику личности которого должен учесть судья, однако предварительный анализ данных конкретного дела мог бы помочь принимать такие решения более объективно, в меньшей степени оставляя их на личное усмотрение судей. Особенно это актуально в контексте принципов ultima ratio и общей интенции уголовно-правовой политики снижать воздействие мер пресечения на обвиняемых и подозреваемых. Однако вопросы, связанные с предотвращением воспроизводства неравенства в уголовном процессе, укладываются в целом в другой комплекс проблемных аспектов, связанных с возможностью внедрения ИИ в процесс принятия судьями решений[22].

Применение ИИ может приводить к воспроизводству социального неравенства и дискриминации практически в любой сфере, поскольку ИИ часто является лишь общим алгоритмом обработки данных (проблема отчуждения). Например, известны случаи, когда ИИ, работающие с распознаванием лиц, не могли принять фотографию лиц из Азии в силу узкого разреза глаз или чернокожих граждан, поскольку в базе было недостаточно их изображений[23]. Поэтому следует отметить, что создать условия для воспроизводства дискриминационных стереотипов могут не только данные, содержащие в себе стереотип, но также и недостаточные данные. Если данные, собранные для обучения ИИ, не представляют полностью все группы, которые необходимо идентифицировать, это может привести к недопустимым выводам. Например, если система распознавания лиц была обучена в основном на изображениях людей европейской внешности, ее способность правильно распознавать лица людей других рас может значительно ухудшиться.

Рассмотренные ошибки в работе ИИ приводят к следующим негативным последствиям в контексте социального неравенства. Во-первых, неравенство в доступе к государственным услугам или предоставлению льгот. Алгоритмы, базирующиеся на порочных данных, могут использоваться при предоставлении услуг или льгот, таких как более выгодные условия кредитования, медицинских услуг или страхования. Это может приводить к тому, что определенным группам населения будет необоснованно отказано в получении услуг или же они получат их в неполном объеме. Во-вторых, закономерным следствием в социокультурном контексте будет упрочение стереотипов. Использующий некорректные данные ИИ будет неосознанно поддерживать существующие стереотипы, что, в свою очередь, может оказывать долгосрочное воздействие на социальные предубеждения и отношения между различными группами[24].

Еще одним негативным результатом может стать потеря доверия к отдельным машинам или программам, к ИИ в целом и снижение доверия к тем, кто его использует. Если пользователи осознают, что алгоритмы, которыми они пользуются, предвзяты, это может снизить уровень доверия к технологиям ИИ и к коммерческим организациям или государственным органам, их применяющим[25].

---

[20] *Шталь Б. К.* Указ соч. С. 29.

[21] *Villasenor J., Foggo V.* (2020) Artificial Intelligence, Due Process, and Criminal Sentencing. *Michigan State Law Review*, no. 295, pp. 334–336. DOI: 10.17613/48wf-jn82

[22] См., напр.: *Taylor I.* (2023) Justice by Algorithm: The Limits of AI in Criminal Sentencing. *Criminal Justice Ethics*, no. 42 (3), pp. 193–213. DOI: https://doi.org/10.1080/0731129X.2023.2275967

[23] *Zou J., Schiebinger L.* (2018) AI can be sexist and racist — it's time to make it fair. *Nature*, no. 559, pp. 324–325. DOI: 10.54254/2754-1169/23/20230367

[24] *Farahani M., Ghasemi G.* (2024) Artificial Intelligence and Inequality: Challenges and Opportunities. *Qeios*. URL: https://doi.org/10.32388/7HWUZ2 (дата обращения 05.09.2024).

[25] Там же.

Возможные пути решения проблем

Безусловно, самым очевидным с юридической точки зрения способом решения поставленных проблем является установление этических стандартов и требований к программированию или использованию таких ИИ с целью недопущения, в частности, дискриминации и в целом воспроизводства социального неравенства. Однако для того чтобы установить подобные этические стандарты, необходимо понять, как именно возможно их сформулировать и какие существуют способы (реальные или потенциальные) преодоления некорректного функционирования ИИ.

В литературе, посвященной этике ИИ, предлагаются две основные стратегии преодоления указанных проблем. Во-первых, это оценка воздействия ИИ по степени влияния на общество или же оценка воздействия на права человека, которая чаще обсуждается на Западе. Из отечественных юридических аналогов можно было бы привести риск-ориентированный подход. В целом концепция оценки сводится к выявлению возможных рисков для равенства, которые могут реализоваться в ходе функционирования таких машин или программ с целью их превентивного устранения. Аналогичным образом организовано предложение в рамках Евросоюза ввести Оценочный список для проверки надежности искусственного интеллекта (ALTAI)[26]. Однако эта модель акцентирует исследовательскую оптику лишь на возможных ошибках, но не предлагает фундаментального конструктивного решения.

Другой подход предполагает составление универсальных этических принципов, заложенных в код программы, на основе которых ИИ не будет допускать дискриминационных различий, даже несмотря на некорректные данные. В рамках этой концепции объясняется, что ошибки в функционировании ИИ, приводящие к воспроизводству неравенства, — это результат невнимания на стадии разработки проекта или технологии к преследуемым ценностям[27]. Основной пресуппозицией этого подхода является суждение о том, что никакая технология ИИ не может быть ценностно нейтральной[28]. Сходный подход мы можем найти в философии М. Маклюэна, исследовавшего медиа, который он формулирует как “the medium is the message”, то есть любое средство передачи информации само по себе несет некоторое значение[29].

Данный подход, с одной стороны, кажется фундаментальным и рефлексивным, но с другой — он в той же мере представляется не вполне конкретизированным и не решающим проблему адаптации этических принципов к функционированию машин. Аналогичную неточность мы можем также увидеть и в одном из комментариев по поводу ошибок в функционировании ИИ: «мусор на входе — мусор на выходе»[30]. Что именно следует идентифицировать как «мусор на входе», как показал анализ, не вполне ясно. Это может быть и искажение данных, умышленное или неосознаваемое, и нерепрезентативная выборка, и недостаток таковых.

Вышеприведенное рассуждение об ошибках в функционировании ИИ, а также причинах их происхождения проблематизирует саму постановку вопроса о возможности этики ИИ. Если верно, что причиной ошибок в действиях ИИ является некорректный набор данных, и дело только в данных, то вопрос об этике машин может быть решен только посредством более тщательного отбора этих данных, а не на уровне этических принципов. Либо упомянутые этические принципы, например принцип равенства, должны быть формализованы, то есть представлены в форме набора отдельных команд, воспринимаемых машинами, или же данных, которые будут соответствовать, по мнению разработчиков, упомянутому принципу. Поэтому вопрос о неравенстве, создаваемом ИИ, относится к этике лишь частично.

---

[26] *Radclyffe Ch., Ribeiro M., Wortham R.* (2023) The assessment list for trustworthy artificial intelligence: A review and recommendations. *Frontiers in Artificial Intelligence*, pp. 2–3. DOI: https://doi.org/10.3389/frai.2023.1020592

[27] *Cavoukian A.* (2017) Global privacy and security, by design: turning the “privacy vs. security” paradigm on its head. *Health Technology*, vol. 7, pp. 329–333. DOI: https://doi.org/10.1007/s12553-017-0207-1

[28] *Шталь Б. К.* Указ соч. С. 42.

[29] *Маклюэн Г. М.* Понимание медиа: внешние расширения человека / Пер. с англ. В. Николаева. М., Жуковский: КАНОН-пресс-Ц, Кучково поле, 2003. 464 с.

[30] *Citrome L.* (2024) Artificial Intelligence and the Potential for Garbage In, Garbage Out. *Current Medical Research and Opinion*, no. 40 (1), pp. 1–2. DOI: 10.1080/03007995.2023.2286785

Представляется, что успешным примером формализации данных принципов является концепция индекса этичности права, которая разрабатывалась в последние годы в Национальном исследовательском университете «Высшая школа экономики». Концепция индекса этичности предполагает совмещение социологического, технологического и этического подходов для составления системы, которая по принципу логического исчисления высказываний сможет ранжировать нормы законодательства на предмет соответствия этих норм моральным принципам. Использование социологических методов для анализа моральных основ как основных критериев нравственной оценки человеческого поведения позволит провести сравнение этических представлений современного российского общества с содержанием юридических норм, обладающих этическим значением[31].

Правовые инструменты преодоления неравенства

Одним из способов решения поставленных проблем является правовое регулирование ИИ с поправкой на возможные ошибки, которые могут приводить к нарушению прав человека. В последние годы сфера ИИ привлекает интерес субъектов, осуществляющих правовое регулирование, а также юридического сообщества в целом[32]. Так, на национальном и международном уровне составляются акты, нацеленные на установление принципов, связывающих компании при разработке технологии ИИ.

Одним из таких актов является Указ Президента Российской Федерации от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации»[33], утверждающий Национальную стратегию развития искусственного интеллекта на период до 2030 г. В данной концепции, в частности, закрепляются основные принципы развития и использования технологий ИИ. К наиболее важным ее принципам относятся защита прав и свобод человека, в том числе права на труд, и предоставление гражданам возможности получать знания и приобретать навыки для успешной адаптации к условиям цифровой экономики; безопасность, в том числе предупреждение и минимизация рисков возникновения негативных последствий использования технологий ИИ; прозрачность, то есть объяснимость работы ИИ и процесса достижения им результатов; а также защищенность, то есть разграничение ответственности организаций — разработчиков и пользователей технологий искусственного интеллекта исходя из характера и степени причиненного вреда, защита указанных пользователей от противоправного применения технологий ИИ.

В рамках данной стратегии был разработан «Кодекс этики в сфере искусственного интеллекта»[34], закрепляющий этические принципы и стандарты поведения, которыми должны руководствоваться акторы (участники отношений) в сфере ИИ. Действие Кодекса распространяется на отношения, связанные с этическими аспектами создания (проектирования, конструирования, пилотирования), внедрения и использования технологий ИИ. Кодекс постулирует примат человекоориентированных и гуманистических подходов, закрепляя их в качестве основных этических принципов и центральных критериев оценки этичного поведения акторов; при этом в качестве наивысшей ценности должен рассматриваться именно человек, его права и свободы. Отдельно следует упомянуть о том факте, что, по мнению разработчиков и подписантов Кодекса, ответственность за последствия применения ИИ всегда должен нести человек; акторы не должны допускать передачи права ответственного нравственного выбора. Кроме того, акторам рекомендуется осуществлять добросовестное информирование пользователей об их взаимодействии с ИИ, когда это затрагивает вопросы прав человека и критических сфер его жизни, а также обеспечивать возможность прекратить подобное взаимодействие по желанию пользователя.

---

[31] *Виноградов В. А., Ларичев А. А.* Индекс этичности права как прикладной инструмент оценки соотношения права и морали // Право. Журнал Высшей школы экономики. 2022. № 5. С. 4–23.

[32] См.: *Конев С. И.* Этико-правовые проблемы регулирования искусственного интеллекта и робототехники в отечественном и зарубежном праве / С. И. Конев, Б. А. Цокова // Образование. Наука. Научные кадры. 2020. № 4. С. 68–73. DOI: 10.24411/2073-3305-2020-10202. EDN: FJBTLD

[33] О развитии искусственного интеллекта в Российской Федерации: Указ Президента РФ от 10.10.2019 № 490 (ред. от 15.02.2024). Собрание законодательства РФ, 14.10.2019. № 41, ст. 5700.

[34] Кодекс этики в сфере искусственного интеллекта. URL: https://a-ai.ru/wp-content/uploads/2021/10/Кодекс_этики_в_сфере_ИИ_финальный.pdf (дата обращения: 04.10.2024).

Примерами других таких актов декларативного характера являются принципы ИИ от Google. Они подразумевают 7 положительных ориентиров и 4 запрета, которых обязуется придерживаться данная корпорация[35].

Еще одним актом являются Азиломарские принципы от 11.08.2017. Они подразделяются на несколько блоков: проведение исследований, этика и ценности, долгосрочная перспектива. Этот акт составлен в виде «открытого письма», которое может подписать любой желающий. На июль 2024 г. он собрал 5720 подписей[36]. Еще одним актом, который исходит не от государств, а от организации — Института инженеров электрики и электроники (Institute of Electrical and Electronics Engineers), — являются принятые в 2017 г. Рекомендации «Искусственный интеллект: исследования, разработки и регулирование» (Artificial Intelligence: Research, Development and Regulation)[37].

Тем не менее нельзя не отметить того факта, что, строго говоря, перечисленные документы, в том числе Национальная стратегия, разработанная в рамках Указа Президента, не являются обязывающими в юридическом смысле этого слова. На настоящий момент такие акты «мягкого права» функционируют лишь в той мере, в которой разработчики добровольно соблюдают такие принципы. В силу этого важное значение приобретает императивное закрепление данных принципов, например, в форме Федерального закона. Между тем принятие подобного акта в настоящее время представляется весьма затруднительным, поскольку технологии ИИ остаются крайне лабильными, а правоприменительная практика в данной сфере еще не сформировалась. Представляется, что формальное закрепление этических стандартов, которым должен соответствовать ИИ, является делом недалекого будущего.

Заключение

В итоге нашего рассуждения мы приходим к выводу о том, что влияние ИИ на социальные взаимоотношения является амбивалентным. Оптика, предлагаемая социальной философией, позволяет увидеть в появлении ИИ не только благо, создающее больше свободного времени, но и инструмент, который может стать причиной нарушения прав человека, увеличения социального расслоения как на уровне отдельных обществ, так и в мировом масштабе.

Мы установили, что в настоящий момент существует внушительная практика применения ИИ в различных сферах права, которая показывает, что функционирование ИИ далеко не всегда приводит к желаемым результатам, а иногда становится причиной дискриминации и воспроизводства социального неравенства.

Проанализировав концепции этики ИИ, мы пришли к выводу, что основной проблемой появления социального неравенства является не столько отсутствие этических принципов в контексте программ, действующих на основе ИИ, сколько недостаточная формализация самих этических принципов на машинном языке. Одним из успешных примеров возможной конвергенции социальной сферы, этики и права представляется разработанный в последние годы индекс этичности права.

Это же замечание релевантно и в отношении правового регулирования данной сферы. Так, существующие акты имеют, как правило, рекомендательный характер, вследствие чего представляется необходимым в ближайшем будущем, по мере замедления бурного развития технологий в сфере ИИ, а также формирования соответствующей правоприменительной практики, формально закрепить этические стандарты, которым должны следовать разработчики ИИ.

Литература/References

1. *Апостолова Н. Н.* Ответственность за вред, причиненный искусственным интеллектом // Северо-Кавказский юридический вестник. 2021. № 1. С. 112–119. DOI: 10.22394/2074-7306-2021-1-1-112-119 (Apostolova, N. N. (2021) Liability for damage caused by Artificial Intelligence. *North Caucasus Legal Vestnik*. Pp. 112–119.)

[35] AI at Google: our principles. URL: https://blog.google/technology/ai/ai-principles/ (дата обращения: 23.07.2024).

[36] Asilomar AI Principles. *Future of Life Institute*. URL: https://futureoflife.org/open-letter/ai-principles/ (дата обращения: 23.07.2024).

[37] Institute of Electrical and Electronics Engineers. IEEE Global Public Policy. URL: https://globalpolicy.ieee.org/wp-content/uploads/2017/10/IEEE17003.pdf (дата обращения: 23.07.2024).

2. *Виноградов В. А., Ларичев А. А.* Индекс этичности права как прикладной инструмент оценки соотношения права и морали // Право. Журнал Высшей школы экономики. 2022. № 5. С. 4–23. (Vinogradov, V., Larichev, A. (2022) The Index of Ethicality of the Law as an applied tool for assessing the correlation between law and morality. *Law Journal of the Higher School of Economics*, no. 5, pp. 4–23. DOI: https://doi.org/10.17323/2072-8166.2022.5.4.23)
3. *Конев С. И.* Этико-правовые проблемы регулирования искусственного интеллекта и робототехники в отечественном и зарубежном праве / С. И. Конев, Б. А. Цокова // Образование. Наука. Научные кадры. 2020. № 4. С. 68–73. DOI: 10.24411/2073-3305-2020-10202. EDN: FJBTLD (Konev, S. I., Tsokova B. A. (2020) Ethical and legal problems of regulating Artificial Intelligence and robotics in domestic and foreign law. *Education. Science. Scientific personnel*, no. 4, pp. 68–73. DOI: 10.24411/2073-3305-2020-10202).
4. *Маклюэн Г. М.* Понимание медиа: внешние расширения человека / Пер. с англ. В. Николаева. М., Жуковский: КАНОН-пресс-Ц, Кучково поле, 2003. 464 с. (McLuhan, M. (1994) Understanding Media: The Extensions of Man. 389 p.)
5. *Маркс К.* Капитал. Т. 1. М.: Государственное издательство политической литературы, 1952. 794 с.
6. *Маркс К.* Экономическо-философские рукописи 1844 г. // К. Маркс, Ф. Энгельс. Собр. соч. Т. 42. М.: Политиздат, 1974. С. 41–174. (Marx, K., Engels, F. (2023) Economic and Philosophic Manuscripts of 1844. Wilder Publications; Edition Unstated. 147 p.)
7. *Шталь Б. К. [и др.]* Этика искусственного интеллекта: кейсы и варианты решения этических проблем / Б. К. Шталь, Д. Шредер, Р. Родригес; пер. с англ. И. Кушнаревой; под науч. ред. А. Павлова; Нац. исслед. ун-т «Высшая школа экономики». М.: Изд. дом Высшей школы экономики, 2024. 200 с. (Stahl, B. C., Schroeder, D., Rodrigues, R. (2023) Ethics of Artificial Intelligence. Case Studies and Options for Addressing Ethical Challenges. *Springer*. DOI: 10.17323/978-5-7598-2981-2).
8. *Юницкий А. Э.* Искусственный интеллект и отчуждение человека от разума: причины, механизмы, последствия / А. Э. Юницкий, Е. О. Петров // Безракетная индустриализация ближнего космоса: проблемы, идеи, проекты: материалы V международной научно-технической конференции, Марьина Горка, 23–24 сентября 2022 г. Минск: ГП «СтройМедиаПроект», 2023. С. 270–281. EDN: MCIVZU. (Unitsky, A., Petrov, E. (2023) Artificial Intelligence and Human Alienation from Mind: Causes, Mechanisms, Consequences. *Collection of Articles of the V International Scientific and Tachnical Conference "Non-Rocket Near Space Industrialization: Problems, Ideas, Projects"*. Pp. 270–281)
9. Bentley, S. V., Naughtin, C. K., McGrath, M. J. [et al.] (2024) The digital divide in action: how experiences of digital technology shape future relationships with artificial intelligence. *AI and Ethics*, no. 3, pp. 1–15. DOI: https://doi.org/10.1007/s43681-024-00452-3
10. Boden, M. [et al.] (2017) Principles of robotics: Regulating robots in the real world. *Connection Sci*, vol. 29 (2), pp. 124–129. DOI: https://doi.org/10.1080/09540091.2016.1271400
11. Cavoukian, A. (2017) Global privacy and security, by design: turning the "privacy vs. security" paradigm on its head. *Health Technology*, vol. 7, pp. 329–333. DOI: https://doi.org/10.1007/s12553-017-0207-1
12. Citrome, L. (2024) Artificial Intelligence and the Potential for Garbage In, Garbage Out. *Current Medical Research and Opinion*, no. 40 (1), pp. 1–2. DOI: 10.1080/03007995.2023.2286785
13. Farahani, M., Ghasemi, G. (2024) Artificial Intelligence and Inequality: Challenges and Opportunities. *Qeios*. DOI: https://doi.org/10.32388/7HWUZ2
14. Gramsci, A. (1974) Selections from the prison notebooks of Antonio Gramsci. *International Publishers, Political Science*. 483 p.
15. Hung, T. W., Yen, C. P. (2023) Predictive policing and algorithmic fairness. *Synthese*, pp. 201–206. DOI: https://doi.org/10.1007/s11229-023-04189-0
16. Mahadi, N., Bakari, A. H. A., Baskaran, S., Maideen, M. H. (2022) How The Pandemic Has Disrupted and Changed Hiring. *International Journal of Academic Research in Business and Social Sciences*, no. 12 (10), pp. 1331–1340. DOI:10.6007/IJARBSS/v12-i10/14825

17. Mugari, I., Obioha, E. (2021) Predictive Policing and Crime Control in the United States of America and Europe: Trends in a Decade of Research and the Future of Predictive Policing. *Social Sciences*, no. 10 (234), pp. 1–14. DOI: 10.3390/socsci10060234
18. Murphy, R., Woods, D. (2009) Beyond Asimov: The Three Laws of Responsible Robotics. Intelligent Systems, IEEE, no. 24 (4), pp. 14–20. DOI: 10.1109/MIS.2009.69
19. Radclyffe, Ch., Ribeiro, M., Wortham, R. (2023) The assessment list for trustworthy artificial intelligence: A review and recommendations. *Frontiers in Artificial Intelligence*, pp. 1–10. DOI: https://doi.org/10.3389/frai.2023.1020592
20. Taylor, I. (2023) Justice by Algorithm: The Limits of AI in Criminal Sentencing. *Criminal Justice Ethics*, no. 42 (3), pp. 193–213. DOI: https://doi.org/10.1080/0731129X.2023.2275967
21. Villasenor, J., Foggo, V. (2020) Artificial Intelligence, Due Process, and Criminal Sentencing. *Michigan State Law Review*, no. 295, pp. 295–354. DOI: 10.17613/48wf-jn82
22. Wallach, W., Allen, C. (2009) Moral Machines: Teaching Robots Right from Wrong. *Oxford Univ. press*. 288 p.
23. Winfield, A., Michael, K., Pitt, J., Evers, V. (2019) Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems. *Proceedings of the IEEE*, no. 107 (3), pp. 509–517. DOI: 10.1109/JPROC.2019.2900622
24. Xinyu, C. (2023) Gender Bias in Hiring: An Analysis of the Impact of Amazon's Recruiting Algorithm. *Advances in Economics, Management and Political Sciences*, no. 23, pp. 134–140. DOI: https://doi.org/10.54254/2754-1169/23/20230367
25. Zou, J., Schiebinger, L. (2018) AI can be sexist and racist — it's time to make it fair. *Nature*, no. 559, pp. 324–326. DOI: 10.54254/2754-1169/23/20230367

Об авторах:

Рыбин Александр Игоревич, аспирант департамента теории права и сравнительного правоведения факультета права Национального исследовательского университета «Высшая школа экономики» (Москва, Российская Федерация); e-mail: airybin@hse.ru

Чащухин Егор Олегович, аспирант департамента теории права и сравнительного правоведения факультета права Национального исследовательского университета «Высшая школа экономики» (Москва, Российская Федерация); e-mail: eochaschukhin@hse.ru

About the authors:

Aleksandr I. Rybin, postgraduate student of the Department of Theory of Law and Comparative of the Faculty of Law, National Research University Higher School of Economics (Moskow, Russian Federation); e-mail: airybin@hse.ru

Egor O. Chashchukhin, postgraduate student of the Department of Theory of Law and Comparative of the Faculty of Law, National Research University Higher School of Economics (Moskow, Russian Federation); e-mail: eochaschukhin@hse.ru