DOI: 10.14515/monitoring.2025.3.2661

А. В. Резаев, Н. Д. Трегубова

ЗАПРЕТИТЬ НЕЛЬЗЯ РЕГУЛИРОВАТЬ: К ВОПРОСУ О ПРОБЛЕМАХ РЕГУЛЯЦИИ ИСПОЛЬЗОВАНИЯ ИИ В ПОВСЕДНЕВНОЙ ЖИЗНИ ОБЩЕСТВА

Правильная ссылка на статью:
Резаев А. В., Трегубова Н. Д. Запретить нельзя регулировать: к вопросу о проблемах регуляции использования ИИ в повседневной жизни общества // Мониторинг общественного мнения: экономические и социальные перемены. 2025. № 3. С. 294—311. https://www.doi.org/10.14515/monitoring.2025.3.2661.

For citation:
Rezaev A. V., Tregubova N. D. (2025) Prohibit Cannot Regulate: On the Question of Adjusting the Use of AI in Everyday Life of Society. *Monitoring of Public Opinion: Economic and Social Changes.* No. 3. P. 294–311. https://www.doi.org/10.14515/monitoring.2025.3.2661. (In Russ.)

Получено: 23.07.2024. Принято к публикации: 25.04.2025.

DOI: 10.14515/monitoring.2025.3.2661

ЗАПРЕТИТЬ НЕЛЬЗЯ РЕГУЛИРОВАТЬ: К ВОПРОСУ О ПРОБЛЕМАХ РЕГУЛЯЦИИ ИСПОЛЬЗОВАНИЯ ИИ В ПОВСЕДНЕВНОЙ ЖИЗНИ ОБЩЕСТВА

PROHIBIT CANNOT REGULATE: ON THE QUESTION OF ADJUSTING THE USE OF AI IN EVERYDAY LIFE OF SOCIETY

РЕЗАЕВ Андрей Владимирович — доктор философских наук, профессор кафедры философии, Ташкентский государственный экономический университет, Ташкент, Узбекистан
E-MAIL: rezaev@hotmail.com
https://orcid.org/0000-0002-3918-835X

ТРЕГУБОВА Наталья Дамировна — кандидат социологических наук, доцент кафедры сравнительной социологии, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия
E-MAIL: n.tregubova@spbu.ru
https://orcid.org/0000-0003-3259-5566

Andrey V. REZAEV[1] — Prof. Dr. habil., Professor of the Department of Philosophy
E-MAIL: rezaev@hotmail.com
https://orcid.org/0000-0002-3918-835X

Natalia D. TREGUBOVA[2] — Cand. Sci. (Soc.), Associate Professor, Chair of Comparative Sociology
E-MAIL: n.tregubova@spbu.ru
https://orcid.org/0000-0003-3259-5566

[1] Tashkent State Economic University, Tashkent, Uzbekistan
[2] St Petersburg State University, St Petersburg, Russia

Аннотация. Статья представляет размышления о том, в какой системе координат может и должно осуществляться регулирование инструментов искусственного интеллекта (ИИ), которые становятся частью повседневной жизни общества. Работа ориентирована на поиск ответа на вопрос: где следует ставить запятую в предложении «запретить нельзя регулировать» в отношении технологий ИИ?

Авторы начинают с общей постановки вопроса, переходя к рассмотрению конкретных проблем, связанных с развитием ИИ (проблемы лжи, проблемы автономных агентов, проблемы «второго перехода» и некоторых других), что позволяет определить сложности, с которыми сталкиваются попытки управлять использованием технологий ИИ. После этого авторы обращаются к характеристике институционального контекста регуляции, выделяя трех ключевых игроков: разработчиков, потребителей и тех, кто определяет «правила игры», — и ставят

Abstract. This article examines the framework governing artificial intelligence (AI) instruments that are becoming an integral part of society's everyday life. The core question framing authors' reflections is: Should AI technologies be prohibited or regulated from the outset?

The paper begins with a general overview of the central question surrounding the development of artificial intelligence (AI), setting the stage for a deeper exploration of critical issues. It examines specific challenges related to AI, including deceptive practices that may arise in machine learning algorithms, the functionalities and implications of autonomous agents in various sectors, and the concept of the "second transition," which refers to the developing state of AI capabilities and their integration into society. Throughout this discussion, the authors highlight the substantial regulatory challenges that arise from these advancements. Building on this foundation, the authors then analyze the ex-

DOI: 10.14515/monitoring.2025.3.2661

вопрос о том, какую роль социальные ученые способы играть в деле регуляции ИИ. В заключение авторы формулируют четыре принципа регуляции: 1) следует регулировать не развитие ИИ, а взаимоотношения между людьми, которые изобретают, внедряют и применяют ИИ; 2) следует регулировать не развитие ИИ, а отрицательные эффекты применения инструментов ИИ в повседневности; 3) на первом этапе регуляция должна ориентироваться на развитие инструментов ИИ в конкретных областях, где выгода от замены человека ИИ наиболее очевидна; 4) существует принципиальная разница между регуляцией ИИ в областях, которые соотносятся с экзистенциальными вопросами (где уже сейчас необходимо вводить безусловные запреты), и в областях, которые с ними не соотносятся (где нужно сначала регулировать, а потом — запрещать).

Ключевые слова: искусственный интеллект, проблема регуляции, взаимозависимость «человек — алгоритм», формальные и неформальные институты, риски и табу

isting regulatory context in detail, emphasizing the distinct roles that developers, consumers, and regulators play in the oversight of AI technologies. They discuss the responsibilities of developers to ensure ethical practices in AI design, the need for informed consumers to navigate complex AI systems, and the imperative for regulators to establish effective frameworks that keep pace with rapid technological changes. Additionally, the paper explores the valuable contributions that social scientists can make in shaping AI regulations by bringing insights into social behavior, ethical considerations, and public policy. This collaborative approach is essential for developing robust regulatory mechanisms that address both the opportunities and challenges of AI integration into society. In conclusion, the authors outline four regulatory principles: 1) The key is not to regulate AI's design and manufacture but the relationships among those who create, deploy, and utilize it; 2) The focus of AI regulation should not be on its making but on addressing the negative effects of using AI tools in everyday life; 3) In the first stage, regulations should focus on the development of AI tools in specific areas where the benefits of replacing humans with AI are most apparent; 4) There is a fundamental difference between regulating AI in areas that relate to existential issues (where we need to introduce prohibitions right now) and other areas (where we should first regulate and then prohibit).

Keywords: artificial intelligence, problem of regulation, human — algorithm interdependence, formal and informal institutions, risks and taboo zones

DOI: 10.14515/monitoring.2025.3.2661

Введение

Цель настоящих размышлений — наметить систему координат для понимания проблемы регуляции технологий искусственного интеллекта (ИИ) в обществе. В своих рассуждениях мы отталкиваемся от вопроса: по аналогии с «казнить нельзя помиловать» где следует поставить запятую в предложении «запретить нельзя регулировать»[1] в отношении технологий ИИ?

Главная проблема с регулированием искусственного интеллекта состоит в следующем: *как можно регулировать нечто, когда люди до конца не понимают, как это нечто работает?*

Исследования и разработки в области искусственного интеллекта (ИИ) начались в середине XX века как проект по созданию вычислительных машин, способных решать задачи, которые умеет решать человек [Russel, Norvig, 2016]. На первом этапе своего развития ИИ был направлен на решение узких задач в контролируемых условиях. Теперь же технологии ИИ становятся частью повседневных практик современного общества. Они являются активными посредниками и участниками социальных взаимодействий и распространились даже на те задачи, которые человек решать не умеет [От искусственного интеллекта…, 2020].

Человека окружает искусственная среда, созданная им самим: города, дороги, дома, даже питание имеет не столько природную, сколько искусственную основу. Все, что создается человеком, не предполагает непреодолимых проблем с регулированием своего сосуществования с человеком и другими обитателями животного мира. Создавая те или иные продукты, человек закладывает в них элементы регуляции: лампочки перегорят и не выдержат высокого напряжения, шины автомобиля не выдержат сверхскоростей и т. д. И сегодня в связи с ускоренным развитием технологий ИИ перед нами встает такой же вопрос: какие ограничения в них закладывать?

Однако проблема регуляции не сводится к техническим ограничениям. Любая совместная жизнь людей регулируется юридическими и моральными установлениями, которые позволяют им сосуществовать. Рассмотрим пример с автомобилями: можно ли разрешать водителям самим принимать решение, по какой стороне улицы ехать и как проезжать перекресток? Риторический вопрос. Соответственно, и развитие технологий ИИ ставит перед нами вопрос о введении юридических правил и этических ограничений, регулирующих их применение в обществе [Etzioni, Etzioni, 2017].

Как именно должна быть организована эта регуляция и на каком этапе развития технологий ее следует вводить? Данные вопросы активно обсуждаются по всему миру. 21 мая 2024 г. Европейский союз после нескольких лет дебатов наконец принял Акт о регуляции технологий ИИ[2]. В Китае, США, Великобритании, России и многих других странах есть стратегии, этические кодексы и отдельные законодательные акты, регулирующие те или иные стороны разработки, внедрения и ис-

[1] Разумеется, мы отдаем себе отчет в том, что запрет — это крайняя степень регуляции. Ставя вопрос подобным образом — запретить *или* регулировать, мы имеем в виду принципиальное решение: стоит ли нечто допускать и вводить правила для его функционирования в обществе либо это нечто следует запретить полностью. См. также обсуждение в нашем Меморандуме: Резаев А. В., Трегубова Н. Д. Человек и/или ChatGPT // Социодиггер. Т. 4. № 5—6. URL: https://sociodigger.ru/articles/articles-page/zapretit-nelzja-regulirovat (дата обращения: 02.06.2025).

[2] Browne R. World's First Major Law for Artificial Intelligence Gets Final EU Green Light // CNBC. 2024. May 21. URL: https://www.cnbc.com/2024/05/21/worlds-first-major-law-for-artificial-intelligence-gets-final-eu-green-light.html (дата обращения: 02.06.2025).

DOI: 10.14515/monitoring.2025.3.2661

пользования технологий ИИ[3]. Исследователи, анализируя принятые стратегии и подходы, фиксируют сходства и различия между ними: общие технологические проблемы, связанные с развитием ИИ, по-разному преломляются и осмысляются в разных национальных и культурных контекстах [Cath et al., 2018; Bareis, Katzenbach, 2022]. Причем ученые и эксперты сходятся в том, что переход от абстрактных юридических и этических положений к конкретным механизмам регуляции требует усилий, которые только начинают предприниматься, и только практика раскроет сильные и слабые стороны нормативных документов.

Предварительные замечания

Начнем наше рассуждение с трех замечаний о регуляции технологий в целом и регуляции ИИ в частности.

Первое. Всякое регулирование начинается *после* того, когда случится нечто плохое[4]. Мы можем предполагать: для того чтобы те, кто будет регулировать ИИ, отнеслись к проблеме серьезно, должен произойти какой-либо существенный кризис, может быть, трагедия.

Второе. Проблема безопасного использования технологий ИИ возникла далеко не вчера[5].

Третье. Не следует рассматривать ИИ как «продукт», нечто подобное самолету или автомобилю, телефону. У ИИ есть своеобразная «самочинность», «самостоятельность», чего нет ни у каких других продуктов человеческой деятельности [От искусственного интеллекта…, 2020]. Инструменты ИИ могут самообучаться, взаимодействовать с другими машинами, принимать функциональные решения без участия человека, выступать самостоятельными агентами при производстве продуктов, связанных с написанием текстов и изображений[6].

Как следствие, с развитием технологий ИИ появляются вопросы, ответов на которые у нас пока нет. Для понимания, с одной стороны, необходимости, а с другой — проблематичности регуляции создания и использования инструментов ИИ рассмотрим некоторые из наиболее очевидных проблем, связанных с ними.

---

[3] См. краткий обзор подобных попыток: Мамедьяров З. А. Регулирование искусственного интеллекта: первые шаги // НИУ ВШЭ. URL: https://foresight.hse.ru/mirror/pubs/share/886272540.pdf (дата обращения: 02.06.2025). Подборка существующих регуляций представлена на сайте AI Ethicist, URL: https://www.aiethicist.org/frameworks-guidelines-toolkits (дата обращения: 02.06.2025). Подробнее с регуляцией ИИ в России можно ознакомиться на специальном сайте: URL: https://ai.gov.ru/ai/regulatory/ (дата обращения: 02.06.2025).

[4] К примеру, когда Генри Форд изобрел и начал продавать свою машину Т-8, не было никаких регуляций, точнее, — вся экономика и регуляции были посвящены лошадям. В США до 30 % экономики были связаны с лошадьми. И только после того как в городах стало невозможно ездить и автомобили стали давить людей больше, чем лошади, начали регулировать новые технологии.

[5] Данный тезис представляется очевидным. Приведем лишь один пример. Илон Маск в интервью с Такером Карлсоном рассказал, почему он решил финансировать Open AI. Он встретился с Ларри Пэйджем — одним из основателей компании Google. Пэйдж сказал Маску, что он создаст настоящий Интеллект. Маск утверждал, что надо уделять внимание безопасности, Пэйджа это не интересовало. Поэтому Маск решил вложиться в открытую для всех (не коммерческую, не зарабатывающую деньги) компанию по разработке ИИ — он дал деньги Сэму Альтману для Open AI.

[6] Говоря про «самочинность» агентов ИИ, следует различать две вещи. Первая — это принципиальная возможность ИИ действовать самостоятельно, автономно осуществлять некоторые действия. Вторая — социальная приемлемость подобных действий, «роль» ИИ в системе общественных отношений, и здесь за ИИ агентность может и не признаваться. Он весьма редко упоминается в качестве самостоятельного автора текстов или изображений, вместо этого обычно сообщают, что продукт был создан при помощи ИИ. Однако в реальной практике создания продукта ИИ выступает именно как агент, как действующее лицо.

DOI: 10.14515/monitoring.2025.3.2661

Проблемы использования технологий ИИ

Проблема инструментов ИИ в интернете

С начала XXI века информация, которую можно было найти в интернете, делилась на две неравноправные части: истинную и ложную. Сначала истинная информация преобладала, затем явно стала преобладать ложная информация. В последнее время, особенно в период пандемии COVID-19 и после ее завершения, люди в подавляющем большинстве используют интернет как источник для получения информации. В ситуации, когда инструменты ИИ имеют очень широкие возможности производить, воспроизводить и навязывать информацию пользователям интернета, те, кто получает информацию из интернета, не то что не застрахованы от ложной информации, у них просто минимизируются шансы получить истинную, никем не контролируемую, никем и ничем не редактируемую информацию.

Еще более насущный пример — дезинформация. Решить проблему дезинформации очень трудно. В первую очередь, потому что коды, позволяющие создавать программы по дезинформации, находятся в открытом доступе и бесплатны. К тому же современные языковые модели (LLMs) (которые уже перестали быть только языковыми, а стали как минимум и визуальными) в состоянии сами разрабатывать коды для дезинформации. Попытки решить данную проблему идут по пути использования одних технологий ИИ для проверки информации, созданной другими технологиями ИИ (см., например, [Lucas et al., 2023; Jiang, Tan, Nirmal, Liu, 2024]). Одно из затруднений данного подхода заключается в том, что различные LLMs лучше всего «нейтрализуются» специализированными программами, которые должны меняться вместе с выходом новых моделей. Данная проблема не представляется полностью непреодолимой технически (см. [Jiang, Zhao, Tan, Liu, 2024]). Вместе с тем для производства дезинформации не требуется особых усилий (LLMs «галлюцинируют» не по задумке разработчиков), в то время как для проверки информации такие усилия требуются. В данной ситуации должны быть очень сильные (финансовые или политические) стимулы для раскрытия дезинформации, чтобы проблема была решена.

Интернет, с одной стороны, стал для многих людей основным источником информации — ненадежным, но дешевым. С другой стороны, инструменты ИИ, удешевляя и делая интернет более доступным[7], одновременно дисквалифицируют, убивают его как источник истинной информации. В поисках истинной информации в интернете люди будут получать в лучшем случае редактируемую в чьих-то интересах информацию или в норме — информацию, которая будет в чем-то истинна, а в чем-то ложна. В худшем случае любая информация из интернета будет ложной, но выглядеть будет как истинная.

По сути дела, инструменты ИИ сводят на нет интернет как источник информации. По всей вероятности, сегодня уже не будет возможности регулировать интер-

[7] Под «удешевлением» интернета мы имеем в виду две вещи. Во-первых, с помощью инструментов ИИ поиск информации становится более «дешевым» с точки зрения затрат времени и усилий. Сначала онлайн-поисковики, такие как Google, а теперь чат-боты, основанные на больших языковых моделях (как ChatGPT), делают поиск информации в интернете простым, доступным каждому. Во-вторых, сама модель бесплатных сайтов и приложений основана на том, что пользователь вместе с безвозмездно оказанной услугой получает постоянный поток рекламной продукции. Причем «бесплатный» интернет не имел бы смысла, не был бы выгоден без постоянной работы онлайн-алгоритмов, включая инструменты ИИ, которые захватывают внимание пользователя, собирают и анализируют данные, на основе предсказаний выдают рекламу/рекомендации. Подробнее см. [Зубофф, 2022; Haidt, 2024].

DOI: 10.14515/monitoring.2025.3.2661

нет в части «отделения зерен от плевел», получения нередактируемой в чьих-либо целях истинной информации. Для решения подобных проблем нужны особые усилия всех стран и народов.

Ложь и ИИ

Ложь является одним из механизмов социальной регуляции. Принципиальная трудность состоит в том, что обучение в обществе (или тренировка и дрессировка) строится на модели «поощрение и наказание». В этом же направлении строится «обучение» машин[8].

Как будет происходить обучение машины машиной? Мы в принципе не можем понимать, как алгоритм реагирует на запрет, на поощрение или наказание. Нам трудно понять это даже в отношении животных. К примеру, как понимает собака или кот, что не надо есть то, что находится на кухонном столе? Животные могут понимать, что нельзя есть со стола, только когда хозяева находятся дома, а не то, что это вообще является «табуированной зоной». Мы не можем понимать и контролировать их процесс «понимания». Отсюда возникает проблема «потери контроля». Мы можем наказать кошку или собаку за неверную интерпретацию, что такое хорошо или плохо, или за неверное исполнение в процессе их обучения нашего желания или команды. За неверную интерпретацию мы можем «наказать» даже льва или медведя, скажем, посадив их в клетку. Но если представить себе, что нашу команду неверно проинтерпретирует более сильное животное, скажем слон или кит? А если нечто еще более сильное? Какова должна быть клетка? Или что-то другое? Следует также учитывать, что, если лев, медведь или другое животное будет достаточно умным, оно потенциально сможет открыть замок клетки.

Подобная же ситуация очевидна и в ситуации с алгоритмами. Недавнее исследование показало, что LLMs способны «обманывать» пользователя, когда при обучении модель выдает желаемый ответ, то есть подстраивается под запрос пользователя, а затем снова возвращается к тому, что было вложено в нее изначально,— авторы называют такое «поведение» *alignment faking* [Greenblatt et al., 2024].

Причем в ситуации с умными алгоритмами мы не можем предвидеть, когда алгоритм станет сообразительным до такой степени, что сам заключит в клетку (тюрьму) тех, кто хочет ограничить его свободу совершать то, что он обучен делать. Может возникнуть ситуация, когда те, кто дал жизнь алгоритму, должны будут опасаться, как бы не быть наказанными за то, что они пытаются ограничить активность своего изобретения. И здесь нет ничего фантастического или принципиально невозможного. К примеру, уже сейчас повсеместна ситуация, когда новый пользователь хочет войти в ту или иную компьютерную систему, а машина просит доказать, что он является человеком, а не машиной, и просит воспроизвести те или иные символы, появляющиеся на экране. И если человек не сможет воспроизвести символы правильно (по мнению машины), ему будет закрыт доступ туда, куда он хочет войти, другими словами, он будет наказан.

[8] См. также рассуждения о неизбежности «искусственных лжецов» [Castelfranchi, 2000].

DOI: 10.14515/monitoring.2025.3.2661

Проблема распространения (proliferation problem)

Еще один вопрос заключается в том, как регулировать распространение исследований и производства ИИ. Распространение ИИ (то есть его вхождение в повседневную жизнь общества), как и все технологические новшества, имеет обоюдоострую составляющую. Классический пример: технология распознавания лиц изначально не предполагала использования для преследования инакомыслящих, а теперь в Китае ее применяют для этого.

Безусловно, уже предпринимались попытки формулировки решений данной проблемы [9]. Однако основное затруднение кроется в создании конкретных социальных механизмов контроля на всем цикле — от разработок до использования ИИ. Проблема здесь не столько в технических решениях, сколько: а) в сочетании разнонаправленных интересов различных заинтересованных сторон, б) в непредвиденных последствиях использования самих инструментов ИИ [10].

Проблема автономных агентов

Возможность и действительность регуляции инструментов ИИ ближайшего будущего — это решение проблемы автономных агентов (Autonomous Agents, AA) [11], которые принимают решения и действуют в режиме реального времени. Мы можем предполагать, что АА, произведенные разными компаниями, на определенном этапе смогут выработать язык для взаимодействия между собой, без возможностей человека регулировать эти взаимодействия.

Суть дела в том, что мы даже не представляем, какие могут возникнуть риски в этой ситуации: чему будут «обучены» эти автономные агенты, как они будут принимать решения в результате взаимодействия с другими автономными агентами. Данной проблемой надо заниматься уже сейчас. Причем временны́е параметры для возникновения автономных агентов — три-пять лет, возможность их взаимодействия на языке, который не знает человек, — еще столько же. Почему так быстро? Потому что во всем мире очень большие вложения в эту область. К тому же это не произойдет одномоментно, а будет постепенный процесс (именно поэтому процесс регуляции инструментов ИИ должен быть постоянным). И маловероятно, что международные институты смогут за это время организоваться для регуляции.

Проблема «второго перехода»

Сегодня ИИ и робототехника остаются раздельными областями исследований и применения. ИИ реализует свою активность, существует в основном в 2D-измерении, в подавляющем большинстве случаев — на компьютерном мониторе. Однако ситуация начинает активно меняться, и инструменты ИИ соединяются с созданными инженерами роботами. ИИ переходит в мир физической реальности 3D.

[9] См., например, полезный обзор по ряду направлений: Наумов В. Б. и др. Правовые аспекты использования искусственного интеллекта: актуальные проблемы и возможные решения (доклад НИУ ВШЭ). URL: https://www.hse.ru/mirror/pubs/share/480106412.pdf (дата обращения: 02.06.2025). Важные принципы и дилеммы регуляции характеризуются также в: [О'Нил, 2018].

[10] Детальный анализ того, как данная проблема проявляется в военной сфере, проведен Генри Киссинджером [Kissinger et al., 2021].

[11] Keary T. Top 5 Autonomous AI Agents You Need to Know About in 2025 // Techopedia. 2023. October 30. URL: https://www.techopedia.com/top-5-autonomous-ai-agents (дата обращения: 02.06.2025).

DOI: 10.14515/monitoring.2025.3.2661

Получается, что вначале — «первый переход» — инструменты ИИ входят в социальную жизнь, начинают взаимодействовать с людьми (Siri, ChatGPT и т. д.), а в реальности «второго перехода» ИИ будет взаимодействовать с машинами, созданными не так, как были созданы сами инструменты ИИ, и уже во взаимодействии с роботами ИИ будет осуществлять активность в обществе. То есть цифровая сущность ИИ соединится с механической сущностью роботов для взаимодействия с людьми и реализации их (людей) целей.

Таким образом, обученные на основе оцифрованной информации и данных в структуре онлайн-реальности инструменты будут приспособляться к логике механических и физических операций для организации взаимодействий с теми, кто не обладает такими основаниями.

Такое положение дел влечет за собой ряд вопросов. Первый вопрос, который с необходимостью возникнет на втором этапе регулирования развития ИИ в эпоху «второго перехода», когда ИИ в том или ином виде будет обладать материализованной в физическом мире оболочкой, состоит в том, чтобы разобраться, сможет ли ИИ «сам» реализовывать варианты своей свободной воли, определять выбор религии, становиться прихожанином той или иной церкви. Еще один вопрос связан с принятием решения об участии ИИ в производстве и хранении, популяризации в музеях продуктов своего творчества. Стоит ли выставлять в художественных музеях картины, созданные ИИ, наравне с произведениями художников-людей, или это должны быть специализированные музеи «произведений» искусства ИИ для людей?

Ключевая проблема «второго перехода», которую можно видеть уже сейчас, состоит в том, что вмешиваться в технические детали регуляции ИИ может только другой ИИ. И здесь мы имеем дело с классической проблемой, сформулированной еще в Древней Греции: кто надзирает за надзирателями?

Институциональный контекст регуляций

Переходя к характеристике институциональных и организационных условий регуляции технологий ИИ, необходимо отметить, что сегодня в деле регуляции технологий ИИ есть три ключевых игрока: разработчики, потребители (общество, люди) и капитализм (правила, по которым живут и разработчики, и потребители). Причем у этих игроков абсолютно разные цели.

Цели разработчиков: а) решать конкретные технические проблемы, б) создавать более мощный ИИ. Цели потребителей — иметь то, что позволит им продолжать жить нормальной жизнью. Например, для сферы высшего образования это означает поступать учиться в учебные заведения, где понятно как, зачем и чему учиться и где потом искать работу, на которой тебя не заменит ИИ.

У капитализма (точнее, у тех, кто разрабатывает правила игры при капитализме) цель очень простая — продолжать разрабатывать правила, по которым будут жить разработчики ИИ и потребители. Если и те, и другие играют по твоим правилам — ты продолжаешь контролировать мир. Главный лозунг здесь: тот, кто *устанавливает правила* для разработки и потребления продуктов ИИ, тот владеет миром. К этой категории относятся и IT-корпорации, и государственные структуры, и международные организации.

DOI: 10.14515/monitoring.2025.3.2661

Сегодня задача состоит в том, чтобы создать социальные институты (то есть формальные правила) и организации, где эти правила будут вырабатывать, и самое главное — эти организации должны определять и контролировать логику и логистику внедрения этих «правил игры», а также их исполнение. Причем очевидно, что при разработке того, как регулировать инструменты ИИ, как определять и обучать ИИ тому, что хорошо или плохо, мы принципиально не можем и не должны полагаться на помощь со стороны ИИ. Данные регулятивные конструкции должны быть разработаны только людьми. Представляется, что уже на данном этапе обеспечить исключительно человеческое участие в разработке и применении того, что может регулировать алгоритмы, крайне сложно. Уже сейчас требуются огромные вложения в организацию исследований, как и в каком направлении решать проблему «потери контроля» над алгоритмами без участия машин и алгоритмов.

Принципиальный момент заключается в том, что организации, определяющие «правила игры», должны быть внешними по отношению к тем учреждениям и компаниям, где разрабатываются и производятся инструменты ИИ. И здесь мы сталкиваемся с фундаментальным противоречием: сегодня теми, кто хоть что-то понимает в сути ИИ, являются именно те организации, которые являются создателями технологий и инструментов ИИ.

В некотором смысле данное противоречие характерно для регуляции любой сферы человеческой деятельности: ученые лучше всего разбираются в научных исследованиях, повара — в приготовлении пищи, производители комбайнов — в комбайнах и т. д. Поэтому любые специалисты будут, и не без основания, претендовать на то, что знают лучше, как и что регулировать в своей области. Вместе с тем в любой области деятельности возникает необходимость *внешнего* контроля, который ограничивал бы действия специалистов, направляя их в соответствии с общественными интересами[12].

В случае технологий ИИ данное противоречие принимает особенно острый характер: для производителей инструментов ИИ — тех, кто разбирается в них лучше всего, — приоритетом будет извлечение прибыли, что может иметь неочевидные средне- и долгосрочные последствия для отдельных потребителей и общества в целом[13].

Что здесь может быть сделано?

Во-первых, государство может более или менее жестко определять «правила игры» и расставлять приоритеты в области создания ИИ и контролировать их соблюдение.

Во-вторых, на нынешнем этапе регулирования принципиальным становится объединение интеллектуальных ресурсов университетов и производителей продуктов ИИ. Университеты намного отстают от организаций-разработчиков по ресурсам и возможностям. У университетов нет ни экономических ресурсов, ни технологи-

[12] Например, в случае Российского научного фонда и те, кто подает заявки, и те, кто их оценивает, — сами ученые. Однако государство обеспечивает регуляцию с помощью определения приоритетных направлений исследования, которым должна соответствовать тематика проектов.

[13] На примере бесконтрольного развития интернета такие последствия убедительно проанализированы в [Haidt, 2024].

DOI: 10.14515/monitoring.2025.3.2661

ческих возможностей стать такой организацией. Однако у них есть интеллектуальные ресурсы, которые должны быть привлечены к работе.

Наконец, к обсуждению проблем ИИ необходимо привлекать лидеров мнений не только из числа ученых: религиозных деятелей, писателей, художников, философов.

Все эти люди — разработчики, представители государства, университетов, религиозных организаций и т. д. — могли бы объединиться в новой организационной структуре, нацеленной на выработку формальных правил внедрения ИИ в повседневную жизнь общества [14].

Кроме базового противоречия следует выделить еще три проблемы, которые возникают при разработке «правил игры» для регуляции технологий ИИ.

Первая проблема — универсальная для любых попыток регулировать социальные процессы: наложение новых правил на старые. Регуляция никогда не происходит «с нуля». В текущей ситуации правила для общественной жизни, выработанные капиталистическим порядком производства и потребления, приходят в резонанс с правилами, которыми руководствуются в своей деятельности алгоритмы ИИ. И эту сложную, плохо просчитываемую в своих эффектах совокупность правил должны регулировать *новые* этические и юридические правила о разработке и использовании ИИ.

Вторая проблема — проблема поиска рациональных оснований для регуляции. Примечательно, что даже наиболее рациональные «айтишники» в отношении регуляции ИИ часто высказываются как визионеры. Они полагаются на интуицию, на собственную (субъективную) веру в то или иное развитие событий, в потенциал технологий ИИ и/или человечества. Страны и наднациональные образования в своих планах и регуляциях руководствуются количественными показателями, которые выглядят рационально, но в действительности очень часто выбраны произвольно либо связаны с краткосрочными тенденциями экономического и политического развития страны. Кроме того, административный аппарат государств ориентируется на общественное мнение своих граждан, которое складывается под влиянием не всегда осознаваемых надежд, страхов и опасений. Таким образом, определение *рациональных* оснований регуляции ИИ чаще всего остается благим пожеланием.

Наконец, третья проблема состоит в противоречии между национальным характером регуляции и глобальным характером влияния технологий ИИ на человечество. Представляется, что в определенный момент понадобятся общие стандарты для оценки деятельности технологий ИИ [Резаев, Трегубова, 2023]. Однако не ясно, на каких основаниях мы могли бы к ним прийти. Пока страны движутся в разных направлениях:

— более строгие ограничения — ЕС;
— контроль технологий и управление обществом — Китай;
— предоставление почти полной свободы технологического развития — Индия;
— минимальная регуляция с ориентацией на права человека — США.

Возможны и промежуточные варианты, как это происходит в современной России.

[14] Возникает вопрос: как мотивировать разработчиков ИИ к участию в создании правил, которые будут ограничивать их прибыль? Ответ очень простой: компании должны понимать, что такие правила будут сформулированы и введены — с их участием или без него.

DOI: 10.14515/monitoring.2025.3.2661

Международные организации также активно высказываются в отношении вопросов, связанных с ИИ. Вместе с тем имеющиеся попытки формулирования принципов и правил регуляции носят весьма общий характер и не являются обязательными к исполнению[15]. Сегодня регуляция ИИ — зона ответственности национальных государств, каждое из которых принимает решения в соответствии с собственными интересами. Поэтому благие пожелания вроде создания «надежного» ИИ или уменьшения разрыва в развитии ИИ-технологий между странами, пока они декларируются на уровне документов международных организаций, останутся лишь пожеланиями[16].

Задачи для социальных наук

Какова роль социальных наук в данной ситуации? Могут ли они предоставить единые рациональные основания для регуляции ИИ?

Представляется, что они как минимум могут попытаться убрать нерациональные основания, выделить их уязвимые места. А также указать на необходимость определить, из какого именно понимания блага людей мы исходим в своем стремлении регулировать ИИ. Можем ли мы договориться о минимальном едином понимании такого блага хотя бы для большинства стран — это отдельный эмпирический вопрос.

В целом же совместимость технологий ИИ с человеком [Russell, 2019] означает не что иное как их *социальную приемлемость*. Строго говоря, сегодня для науки изучать ИИ — значит изучать не то, как работает человеческий мозг, а то, какова взаимозависимость человека и ИИ в разных культурах, различных биологических и социальных средах.

Сущность происходящих изменений (положительных и отрицательных), которые зависят от внедрения технологий ИИ, имеет абсолютно социальный характер. Они относятся к социальной сфере, к общим человеческим условиям совместной жизни. Прозвучавший в марте 2023 г. призыв лидеров и ведущих разработчиков современных инструментов ИИ[17] «сделать паузу», приостановиться и оглядеться не был услышан именно теми, кем этот призыв должен был быть услышан в первую очередь, — социальными учеными, политиками, людьми, принимающими решения в отношении социальной и гуманитарной проблематики развития ИИ.

В нашей ситуации проблема социальных ученых состоит в том, что они реагируют на развитие ИИ постфактум, не участвуя в разработке технологий и принятии нормативных документов. Задачи социальных наук лежат в плоскости определения того, как минимизировать риски и предотвратить ущерб от вхождения ИИ в повседневную жизнь людей.

---

[15] В качестве примера можно привести резолюцию ООН о регуляции ИИ: Hickman T. AI Watch: Global Regulatory Tracker — United Nations // White & Case. 2024. May 13. URL: https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-united-nations (дата обращения: 02.06.2025).

[16] Отметим, что и сами представители международных организаций это понимают. Так, заместитель директора МВФ Гита Гопинат, характеризуя роль международных финансовых организаций в определении общих оснований для регулирования ИИ, сводит ее к а) накоплению знаний об использовании ИИ в глобальном масштабе, б) организации форумов для обмена опытом, в) консультированию. Активная роль остается за государствами. См.: Гопинат Г. Использование ИИ на благо всего мира // Международный валютный фонд. 2023. Декабрь. URL: https://www.imf.org/ru/Publications/fandd/issues/2023/12/ST-harnessing-AI-for-global-good-Gita-Gopinath (дата обращения: 02.06.2025).

[17] Pause Giant AI Experiments: An Open Letter // Future of Life Institute. 2023. URL: https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (дата обращения: 02.06.2025).

DOI: 10.14515/monitoring.2025.3.2661

На основании каких принципов это могло бы быть сделано?

Пока данный текст готовился к публикации, в литературе, в основном англоязычной, появилось много исследований, аналитических статей и других материалов, посвященных проблемам социальных последствий внедрения ИИ, регуляции и контролю над технологиями ИИ. Особенно активно стали появляться работы по этим проблемам после проведения Саммита по ИИ во Франции[18]. В многоголосье различных предложений по регуляции и попыток их осмысления нетрудно потеряться.

Поэтому здесь мы бы хотели обратить внимание на одну очень важную, как нам представляется, публикацию, появившуюся в апреле 2025 г. и подготовленную ведущими специалистами Google по безопасности ИИ. Работа Анки Драган[19] и ее коллег[20] представляет интерес для социологов по нескольким ключевым направлениям, связанным с новой социальной аналитикой, этикой и социальными последствиями внедрения ИИ. В ней специалисты компьютерных наук в полный голос призывают обратить внимание на *риски*, и не только на те, которые уже существуют сегодня, но и на те, которые появятся завтра. Они подразделяют их на четыре категории: злоупотребления со стороны пользователей (misuse), несоответствие ИИ человеческим целям и ценностям (misalignment), ошибки ИИ и, наконец, структурные риски, связанные с взаимодействием алгоритмов в рамках сложных систем.

В то время как сами авторы сосредотачиваются на проблемах первых двух типов, наибольший интерес и наибольшие затруднения вызывает последняя категория. Именно здесь — в рамках анализа плохо предсказуемых взаимодействий людей и агентов ИИ — становится очевидной необходимость координации усилий представителей различных дисциплин для понимания взаимозависимостей «человек — машина» и «машина — машина». Очевидно, что для участия в решении подобных проблем социологам будет необходимо переосмыслить такие категории, как «агентность», «ответственность», «доверие», «управление» и «регулирование».

Риски, угрозы и табу

Итак, вместе с развитием технологий и инструментов ИИ появляются риски и угрозы, с которыми человек и общество в принципе не могли сталкиваться в индустриальную эпоху. Это обстоятельство уже находит свое отражение в нормативной документации. Так, Акт о регуляции ИИ, недавно принятый ЕС, предлагает классификацию технологий ИИ по тому, какие риски (от неприемлемых до минимальных) они предполагают.

Наличие рисков подводит нас к мысли о необходимости безусловных запретов (табуированных зон) в использовании технологий ИИ [Rezaev, 2021]. По мысли

[18] Саммит по ИИ во Франции состоялся 10—11 февраля 2025 г. в Париже. Мероприятие, известное как AI Action Summit, собрало более тысячи участников, включая первых лиц государств, из более чем ста стран. См.: URL: https://www.elysee.fr/en/sommet-pour-l-action-sur-l-ia (дата обращения: 02.06.2025).

[19] Анка Драган (Anca Dragan) — директор по безопасности и согласованию ИИ (Director of AI Safety and Alignment) в Google DeepMind, одновременно — доцент (Associate Professor) кафедры электротехники и компьютерных наук (EECS) Калифорнийского университета в Беркли. См.: Tao G. Anca Dragan named Head of AI Safety and Alignment at Google DeepMind // Berkeley Electrical Engineering & Computer Sciences. 2024. March 28. URL: https://eecs.berkeley.edu/news/anca-dragan-named-head-of-ai-safety-and-alignment-at-google-deepmind/ (дата обращения: 02.06.2025).

[20] Dragan A., Shah R., Flynn F., Legg Sh. Shane Taking a Responsible Path to AGI // Google DeepMind. 2025. April 2. URL: https://deepmind.google/discover/blog/taking-a-responsible-path-to-agi/ (дата обращения: 02.06.2025).

DOI: 10.14515/monitoring.2025.3.2661

Мэри Дуглас [Дуглас, 1994], в современном обществе понятие риска становится одним из ключевых для характеристики существования человека. Если изначально понятие «риск» было связано с исчислением вероятностей, то сегодня в повседневной жизни риск — это прежде всего риск подвергнуться некой угрозе. Важность данного понятия указывает на уязвимый характер существования человека. Дуглас отмечает: если в традиционных обществах соблюдение табу связано с попыткой оградить *общество* от нежелательных последствий действий индивидов, то в современном обществе понятие риска указывает на стремление оградить *индивида* от нежелательного воздействия общества.

Таким образом, «риск» маркирует возможную угрозу для индивида: риски нужно минимизировать, а угрозы — предотвращать. «Табу» (табуированные зоны) — это область безусловных запретов, охраняющая то, что является сакральным в данном обществе, то, без чего социальный порядок разрушится. При этом наличие рисков оказывается легче если не просчитать, то обосновать: угроза конкретному человеку в современном мире более понятна, чем угроза обществу.

Сегодня одна из насущных проблем регуляции технологий ИИ состоит в том, чтобы определить не только риски, но и те табуированные зоны, в которых люди не должны соприкасаться с ИИ.

Таковы общие суждения. Но понимаем ли мы, *что именно* и *как именно* мы должны запрещать? Представляется, что ответ на поставленный вопрос — «нет». И здесь мы переходим к ключевой части нашей аргументации.

Запретить или регулировать?

Вернемся к вопросу, поставленному в начале статьи: где должна стоять запятая во фразе «запретить нельзя регулировать»?

Представляется, что вопрос один, а ответа должно быть два: сначала запятая должна быть после «нельзя», а потом, когда появятся люди, понимающие, что можно и нужно запретить, — запятую нужно будет ставить после слова «запретить». Регуляция, особенно на начальных этапах, должна быть ориентирована не на *запретить*, а на *создать*, *организовать*.

Для *изменения* системы нужно знать эту систему. Как во времена Форда, когда в начале XX века вся экономика США была ориентирована на лошадей (фураж, дороги, снабжение, производство): производство автомобилей не могло изменить только экономику — изменилась вся общественная конструкция. Никто ничего не запрещал *в начале* производства автомобилей. А потом появились люди, которые социализировались в новой системе и начали говорить, что нужно запретить (сначала скорость, а потом другое октановое число бензина, а не наоборот). Запрет — это всегда элемент культуры.

Так и с ИИ: сначала надо регулировать и обучать людей, как жить и работать в новой системе. А затем обученные в новой системе люди скажут, что нужно *запрещать*.

Можно говорить о том, что сегодня регуляция инструментов ИИ идет именно в этом направлении. Создаются своего рода «песочницы» для экспериментального внедрения новых технологий. Так, в России действует закон «Об экспериментальных правовых режимах в сфере цифровых инноваций в Российской

DOI: 10.14515/monitoring.2025.3.2661

Федерации»[21], а конкретно в столице нашей родины — закон «О проведении эксперимента по установлению специального регулирования в целях создания необходимых условий для разработки и внедрения технологий искусственного интеллекта в субъекте Российской Федерации — городе федерального значения Москве и внесении изменений в статьи 6 и 10 Федерального закона „О персональных данных“»[22].

Вместе с тем главное в сегодняшней ситуации заключается в том, чтобы определить *как, кто и чему будет обучать* тех, кто обязательно станет регулировать использование ИИ. Сейчас они обучаются с помощью «социальных медиа», очень отрывистых контактов на различных выставках, форумах, саммитах (где суть дела — в необходимости что-то продать, получить выгоду, а не объяснить или понять), сигналов «сверху» — это в лучшем случае. В худшем случае — не обучаются нигде, а будут действовать «как прикажут».

Четыре принципа регуляции

Сформулируем принципы регуляции инструментов ИИ, которые могут быть применимы на первом этапе — «запретить нельзя, регулировать».

Первый принцип первого этапа регулирования состоит в том, что надо регулировать не развитие ИИ, а *взаимоотношения* между людьми, которые изобретают, внедряют и применяют ИИ в повседневной жизни[23].

Второй принцип: следует регулировать не развитие ИИ, а *отрицательные эффекты* применения инструментов ИИ в повседневности. Соответственно, задача именно социальных ученых — определить систему координат, методологию и логику эмпирических исследований, в которых следует фиксировать эти отрицательные эффекты применения ИИ в повседневности.

Третий принцип: регуляция на «первых порах» должна ориентировать развитие инструментов ИИ на то, чтобы найти им применение *в трех фундаментальных направлениях*: инструменты ИИ должны быть активно использованы:

а) в опасных для жизни человека областях;

б) в рутинизированных и повторяющихся, нетворческих областях;

в) в дизайне, производстве инструментов ИИ, находящихся в открытом доступе и не являющихся патентованным продуктом больших компаний-производителей, таких как Google, Netflix, Meta[24].

Наконец, *четвертый принцип* заключается в том, что существует принципиальная разница между регуляцией в областях, которые соотносятся с экзистенциаль-

[21] Федеральный закон от 31.07.2020 № 258-ФЗ (ред. от 08.08.2024) «Об экспериментальных правовых режимах в сфере цифровых инноваций в Российской Федерации» (с изм. и доп., вступ. в силу с 05.01.2025) // Консультант. URL: https://www.consultant.ru/document/cons_doc_LAW_358738/314a79f49a806b40b413fda2b160c63163cb3d6d/ (дата обращения: 02.06.2025).

[22] Федеральный закон «О проведении эксперимента по установлению специального регулирования в целях создания необходимых условий для разработки и внедрения технологий искусственного интеллекта в субъекте Российской Федерации — городе федерального значения Москве и внесении изменений в статьи 6 и 10 Федерального закона „О персональных данных“ от 24.04.2020 № 123-ФЗ» (последняя редакция) // Консультант. 2020. 24 апреля. URL: https://www.consultant.ru/document/cons_doc_LAW_351127/ (дата обращения: 02.06.2025).

[23] Данный принцип может быть очевиден для философов, юристов, социологов. Но представляется, он не всегда очевиден для разработчиков или администраторов.

[24] Компания и соцсети, которыми она владеет, признаны в России экстремистскими и запрещены.

DOI: 10.14515/monitoring.2025.3.2661

ными вопросами жизни и смерти человека, и в областях, которые с ними не соотносятся. Несмотря на общее положение о том, что нужно сначала понять, а потом запрещать, в областях первого рода могут быть сразу определены табуированные зоны. Их введение основано на очень простом положении: мы *уже* понимаем, что алгоритмы ИИ не похожи на людей в том, что они не рождаются, не умирают, не радуются и не страдают.

Рассмотрим приложение данных принципов к текущей ситуации развития больших языковых моделей.

«Социальные медиа», которые не регулировались и принесли не меньше вреда, чем пользы (возможно, и больше — зависит от методики измерения) для «социальности» общества, были построены скорее по принципу дать больше выгоды и больше дохода, чем объединить общества/сообщества. По этому же принципу — доход и выгода — функционируют сегодня и компании, разрабатывающие большие языковые модели. Ситуация складывается следующим образом: с 2022 г. происходит объединение потенциалов социальных медиа с инструментами ИИ (языковыми моделями), и оно не может служить на благо развития общества в целом, поскольку будет ориентировано на получение прибыли теми, кто переведет инструменты ИИ в массовое производство, цель (прибыль) будет оправдывать средства и отодвинет человеко-ориентированную модель развития ИИ.

Поэтому регулировать надо в первую очередь этот «союз».

Выводы

Подводя итоги рассуждения, выделим несколько ключевых тезисов.

1) Проблема регуляции ИИ — гораздо более широкая и сложная проблема, чем это представляется юристам и чиновникам, стремящимся к однозначным и фиксированным ее решениям. В эпоху ИИ в деле регулирования не может быть простого дихотомического деления «запретить-разрешить».

2) Регуляция применения ИИ в повседневной жизни — это постоянный процесс, а не одно(несколько)моментные действия запретительного характера.

3) В эпоху ИИ нельзя идти от «положительного» опыта регулирования, который был в прошлом. Слишком быстро происходят изменения, и суть регуляции должна быть в «предвосхищении» будущей ситуации и гибкой реакции на изменения в режиме реального времени.

4) Невозможно регулировать исследования ИИ, но можно и нужно регулировать а) отношения между разработчиками, производителями и пользователями ИИ; б) использование ИИ в определенных сферах, определять запретные (табуированные) зоны.

5) В разных отраслях, очевидно, необходимо будет регулировать разные вещи. Наиболее важное различие пролегает между сферами, где использование инструментов ИИ не допускается в принципе, и остальными сферами, где должны быть постепенно выработаны соответствующие правила и нормы.

6) Важно не допустить монополизации производства инструментов ИИ, поддерживать уровень «соревнования» между компаниями, которые производят ИИ.

7) Для регулирования необходимы «новые люди» — междисциплинарные специалисты, которых еще только предстоит подготовить, а не просто «компьютерщи-

DOI: 10.14515/monitoring.2025.3.2661

ки», «юристы» или «социологи». Их подготовка — одна из ключевых задач на ближайшее будущее.

8) Принципиальные риски будут исходить не только из свойств инструментов ИИ, но и, как это всегда бывает в использовании новых технологий, из человеческой глупости и человеческих страстей.

Список литературы (References)

1. Дуглас М. Риск как судебный механизм // THESIS. 1994. Вып. 5. С. 242—253.
Douglas M. (1994) Risk as a Forensic Resource. *THESIS*. No. 5. P. 242—253. (In Russ.)

2. Зубофф Ш. Эпоха надзорного капитализма. Битва за человеческое будущее на новых рубежах власти. М.: Издательство Института Гайдара, 2022.
Zuboff Sh. (2022) The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. Moscow: Gaidar Institute Press. (In Russ.)

3. О'Нил К. Убийственные большие данные. Как математика превратилась в оружие массового поражения. М.: АСТ, 2018.
O'Neil C. (2018) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Moscow: AST. (In Russ.)

4. От искусственного интеллекта к искусственной социальности: новые исследовательские проблемы современной социальной аналитики / под ред. А. В. Резаева. М.: ВЦИОМ, 2020.
Rezaev A. V. (ed.) (2020) Artificial Intelligence on the Way to Artificial Sociality: New Research Agenda for Social Analytics. Moscow: VCIOM.

5. Резаев, А. В., Трегубова, Н. Д. Возможность и необходимость человеко-Ориентированного искусственного интеллекта в юридической теории и практике // Journal of Digital Technologies and Law. 2023. Т. 1. № 2. С. 564—580. https://doi.org/10.21202/jdtl.2023.24.
Rezaev A. V., Tregubova N. D. (2023) The Possibility and Necessity of the Human-Centered AI in Legal Theory and Practice. *Journal of Digital Technologies and Law*. Vol. 1. No. 2. P. 564—580. https://doi.org/10.21202/jdtl.2023.24.

6. Bareis J., Katzenbach C. (2022) Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics. *Science, Technology, & Human Values*. Vol. 47. No. 5. P. 855—881. https://doi.org/10.1177/01622439211030007.

7. Castelfranchi C. (2000) Artificial liars: Why Computers Will (Necessarily) Deceive Us and Each Other? *Ethics and Information Technology*. Vol. 2. P. 113—119. https://doi.org/10.1023/A:1010025403776.

8. Cath C., Wachter S., Mittelstadt B., Taddeo M., Floridi L. (2018) Artificial Intelligence and the 'Good Society': The US, EU, and UK Approach. *Science and Engeneering Ethics*. Vol. 24. P. 505—528. https://doi.org/10.1007/s11948-017-9901-7.

9. Etzioni A., Etzioni O. (2017) Should Artificial Intelligence Be Regulated? *Issues in Science and Technology*. Vol. 33. No. 4. P. 32—36.

10. Greenblatt R., Denison C., Wright B., Roger F., MacDiarmid M., Marks S., Treutlein J., Belonax T., Chen J., Duvenaud D., Khan A., Michael J., Mindermann S., Perez E., Petrini L., Uesato J., Kaplan J., Shlegeris B., Bowman S. R., Hubinger E. (2024) Alignment Faking in Large Language Models. https://doi.org/10.48550/arXiv.2412.14093.

11. Haidt J. (2024) The Anxious Generation: How the Great Rewiring of Childhood Is Causing an Epidemic of Mental Illness. New York: Penguin Press.

12. Jiang B., Tan Z., Nirmal A., Liu, H. (2024) Disinformation Detection: An Evolving Challenge in the Age of LLMs. In: *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*. P. 427—435. https://doi.org/10.1137/1.9781611978032.50.

13. Jiang B., Zhao C., Tan Z., & Liu H. (2024) Catching Chameleons: Detecting Evolving Disinformation Generated using Large Language Models. In: *Proceedings of 2024 IEEE 6th International Conference on Cognitive Machine Intelligence (CogMI)*. P. 197—206. https://doi.org/10.1109/CogMI62246.2024.00034.

14. Lucas J. S., Uchendu A., Yamashita M., Lee J., Rohatgi S., Lee D. (2023) Fighting Fire with Fire: The Dual Role of LLMs in Crafting and Detecting Elusive Disinformation. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. P. 14279—14305. https://doi.org/10.18653/v1/2023.emnlp-main.883.

15. Kissinger H., Schmidt E., Huttenlocher D. (2021) The Age of AI: And Our Human Future. Boston, MA: Little, Brown and Company.

16. Rezaev A. V. (2021) Twelve Theses on Artificial Intelligence and Artificial Sociality. *Monitoring of Public Opinion: Economic and Social Changes*. No. 1. P. 20—30. https://doi.org/10.14515/monitoring.2021.1.1894.

17. Russell S. (2019) Human Compatible: Artificial Intelligence and the Problem of Control. New York: Viking.

18. Russell S., Norvig P. (2016) Artificial Intelligence: A Modern Approach. Harlow: Pearson Education Limited.