DOI: 10.24412/1994-3776-2024-3-180-186 УДК: 316.4

Д.А. Багдасарян

Оценка и регулирование последствий внедрения искусственного интеллекта в деятельность предприятий и организаций

D. Bagdasaryan. Assessment of negative consequences of the introduction of artificial intelligence and management of emerging social risks

Аннотация. В статье рассматриваются негативные последствия внедрения искусственного интеллекта (ИИ) и методы управления возникающими социальными рисками. Анализируются этические, экономические, социальные, культурные, политические и правовые аспекты воздействия ИИ. Описаны количественные и качественные методы оценки рисков, а также стратегии смягчения последствий, включая переподготовку работников, законодательные инициативы и просветительские кампании. Рассматриваются международные стандарты и примеры регулирования рисков ИИ.

Ключевые слова: искусственный интеллект; социальные риски; риск-менеджмент; регулирование ИИ.

Контактная информация: 193023, Санкт-Петербург, Садовая ул., д.21; тел.: (909) 583-25-46; e-mail: bagdasaryan.di@gmail.com

Abstract. The article considers the negative consequences of the introduction of artificial intelligence (AI) and methods of managing the resulting social risks. Ethical, economic, social, cultural, political and legal aspects of the impact of AI are analysed. Quantitative and qualitative methods of risk assessment are described, as well as mitigation strategies, including worker retraining, legislative initiatives and awareness campaigns. International standards and examples of AI risk management are reviewed.

Keywords: artificial intelligence; social risks; risk management; AI regulation.

Contact information: 21 Sadovaya street, Saint-Petersburg, 193023; тел.: (909) 583-25-46; e-mail: bagdasaryan.di@gmail.com

В настоящее время искусственный интеллект (ИИ) обладает потенциалом для значительного улучшения качества жизни, экономического развития и научных достижений. Однако неконтролируемое внедрение ИИ может привести к серьезным социальным и экономическим последствиям, таким как увеличение безработицы, усиление неравенства, нарушение прав человека и другим проблемам. В связи с этим, мы считаем актуальным проанализировать существующие стратегии и подходы к оценке и минимизации негативных последствий ИИ, а также оценить эффективность текущих методов управления социальными рисками и их потенциал для масштабирования.

Трансформация жизнедеятельности общества, в частности, и повсеместное внедрение ИИ, является причиной роста социальных рисков, опасностей и угроз, которые могут оказать деструктивное воздействие на процессы производства, взаимодействие между людьми, их социальную защищенность и т. д. [1]. В целом все риски можно разделить на три группы, в зависимости от объектов воздействия: техно-экономические (воздействуют на бизнес-структуры, государственные учреждения, органы государственной власти), информационно-технологические (воздействуют на личность, социальные группы, бизнес-структуры, органы государственной власти, институты гражданского общества) и социальные риски (личность, социальные группы, социальные институты общества) [1].

Багдасарян Диана Арцруновна – аспирант кафедры социологии и управления персоналом Санкт-Петербургского государственного экономического университета

D. Bagdasaryan – Post-graduate student of Sociology and Human Resource Management Department of Saint-Petersburg State Economic University

Описанные группы рисков могут привезти к положительным и негативным последствиям, однако в рамках данной статьи мы уделяем внимание только негативным эффектам. Можно выделить четыре группы негативных последствий внедрения ИИ (см. табл. 1). Первая – этические проблемы, например, связанные с дискриминацией и предвзятостью алгоритмов, а также прозрачностью и объяснимостью решений ИИ. Вторая группа – экономические последствия, например, замещение рабочих мест роботами и вызванная этим безработица. Также к этой группе можно отнести возникновение экономического неравенства и неравномерную концентрацию богатства. К третьей группе мы относим социальные и культурные последствия, такие как влияние на социальное взаимодействие и межличностные отношения и культурные изменения и потери традиционных ценностей. К последней, четвертой группе, мы относим политические и правовые последствия внедрения ИИ: вопросы ответственности и подотчетности, а также правовое регулирование ИИ [7].

Таблица 1. Группы негативных последствий внедрения ИИ

Группа негативных последствий	**Содержательные характеристики негативных последствий**
Этические последствия	– социальная и расовая дискриминация – предвзятость алгоритмов – отсутствие прозрачности и объяснимости решений ИИ – нарушение конфиденциальности в процессе сбора и анализа big data – нарушение прав на конфиденциальность во время отслеживания и идентификации людей – потеря контроля и доверия в принятии решений автономных систем – подрыв способности принятия людьми самостоятельных и осознанных решений – уменьшение прямых социальных контактов и усиление социальной изоляции из-за увеличения использования ИИ – вопрос ответственности за ошибки или вред, причиненный ИИ-технологиями
Экономические последствия	– замещение рабочих мест, уменьшение спроса на определенные виды работы и безработица – усиление разрыва между богатыми и бедными – отстранение стран от выгод, связанных с ИИ, из-за ограниченного доступа к технологиям или недостаточной подготовленности – концентрация экономической власти в руках тех, кто владеет и контролирует ИИ – экономические потери из-за кибератак – неопределенность и риски для инвесторов и бизнеса в связи с быстрым развитием ИИ и его внедрением – негативное влияние на экологию из-за большой энергоемкости
Социальные и культурные последствия [2]	– уменьшение личной коммуникации и социальных воздействий лицом к лицу – склонность людей к ограничению социальных взаимодействий – культурная гомогенизация и глобальная ассимиляция вследствие распространения глобальных культурных норм и стандартов – изменение культурных практик, как люди воспринимают, практикуют и передают культурные нормы и традиции – укрепление уже существующих предубеждений и стереотипов в обществе – поляризация общества [12] – манипулирование общественным мнением
Политические и правовые последствия	– создание фейковых новостей и влияние на общественное мнение в политических процессах – использование ИИ для усиления контроля над гражданами, мониторинга их действий и подавления политических оппозиций – способствование введению алгоритмического влияния на принятие политических решений, что может привести к коррупции и несправедливости в распределении ресурсов и власти – юридическая неопределенность

Далее рассмотрим методы оценки негативных последствий внедрения ИИ, которая направлена на выявление потенциальных негативных последствий внедрения ИИ, включая

экономические, социальные, культурные, политические и правовые аспекты. К методам оценки относятся количественные и качественные методы.

Количественные методы оценки включают в себя использование статистических данных и моделей для анализа негативных последствий внедрения ИИ. В данном методе мы выделили два подхода: статистический анализ и эконометрические модели. Подход статистического анализа предполагает три метода работы с информацией: сбор данных, анализ тенденций и прогнозирование.

Качественные методы фокусируются на сборе и анализе качественной информации, которая может дать глубокое понимание социальных и культурных аспектов негативных последствий ИИ. К ним относятся интервью и организация фокус-групп.

На наш взгляд, для наиболее эффективного анализа следует использовать и количественные подходы, и качественные, так как они взаимно дополняют друг друга и позволяют получить более полное представление о негативных последствиях внедрения ИИ.

После того, как был проведен анализ последствий, можно понять масштаб и характер рисков, связанных с внедрением ИИ, что позволяет информировать процесс принятия решений и разработки стратегий управления этими рисками. Для управления ими требуется комплексный подход, включающий стратегии смягчения последствий, регулирование и законодательные инициативы, а также работа с общественным сознанием и образованием (см. табл. 2).

Таблица 2. Способы управления социальными рисками, возникающими при внедрении ИИ

Название способа	Содержательная характеристика
Стратегии смягчения последствий	
Развитие навыков и переподготовка работников	– Разработка и финансирование программ переквалификации и повышения квалификации для работников, чьи профессии станут излишними из-за автоматизации – Введение социальных гарантий, таких как компенсации и поддержка для работников, переходящих на новые рабочие места или претерпевающих профессиональную переориентацию
Создание новых рабочих мест в смежных отраслях	– Поддержка и стимулирование развития новых технологий и отраслей, которые могут создать рабочие места – Направление инвестиций в секторы, требующие человеческого труда и предполагающие высокие интеллектуальные или креативные способности, которые трудно заменить автоматизацией
Регулирование и законодательные инициативы	
Разработка и внедрение этических норм и стандартов	– Разработка этических норм и стандартов для проектирования и использования ИИ, включая принципы справедливости, прозрачности и ответственности – Аудит и сертификация ИИ-систем для проверки соответствия этическим стандартам и соблюдения законодательства
Законодательные меры и международное сотрудничество	– Принятие законодательства, направленного на регулирование использования ИИ, включая защиту данных, приватность и предотвращение дискриминации – Содействие международным стандартам и нормативам в области ИИ для обеспечения консистентности и эффективности правового регулирования
Работа с общественным сознанием и образованием	
Просветительские кампании и повышение общественной осведомленности	– Проведение информационных кампаний для общественности о том, как ИИ влияет на их жизнь, права и свободы – Организация открытых дискуссий и консультаций с общественностью для обсуждения этических и правовых вопросов, связанных с использованием ИИ
Образовательные программы по ИИ и цифровой грамотности	– Введение образовательных программ по ИИ и цифровой грамотности в школьные и университетские курсы – Подготовка специалистов и руководителей с пониманием этических и правовых аспектов ИИ

Рассмотрим примеры в каждой из групп способов управления социальными рисками, возникающими при внедрении ИИ.

Примеры стратегии смягчения последствий внедрения ИИ можно найти в инициативах по переподготовке и поддержке работников. Так, канадская организация Future Skills Centre финансирует и проводит исследования, направленные на понимание будущих потребностей в навыках и разработку программ обучения для работников. Целью программ является подготовка и переквалификация трудовых ресурсов для того, чтобы они могли успешно адаптироваться к изменениям на рынке труда, включая автоматизацию и внедрение новых технологий, включая ИИ.

В России данная стратегия реализуется программами повышения квалификации и формирования новых профессиональных навыков. Благодаря государственной программе «Содействие занятости населения» граждане могут пройти первичное тестирование, получить поддержку психолога, возможность переобучения по востребованным у работодателей профессиям, обрести уверенность в себе, а также навыки самопрезентации. Кроме того, участники программы заключают договор с будущим работодателем, которого подбирают также во время профориентации. В 2023 году новые знания получили более 850 тыс. человек, 80% из них трудоустроены [4].

Регулирование ИИ и законодательные инициативы активно внедряются во всем мире. Так, Международная организация по стандартизации (ISO) разработала два знаковых стандарта: ISO 42001 и ISO / IEC 23894, обновляемых каждый год [9]. ISO 42001 охватывает более широкие вопросы управления ИИ: рекомендации по созданию, внедрению, обслуживанию и совершенствованию системы управления ИИ, в том числе и вопросы риск-менеджмента. В то же время в ISO / IEC 23894 («ИИ – управление рисками») уделяется особое внимание управлению рисками ИИ. В стандарте признается, что, хотя ИИ разделяет некоторые риски с традиционными программными системами (например, уязвимости в системе безопасности), он также создает уникальные риски.

В стандарте ISO / IEC 23894 предложены следующие способы управления рисками (см. табл. 3). Важно отметить, что документ требует от компаний регулярного пересмотра и обновления своих оценок рисков, которые будут появляться в большем объеме.

Таблица 3. Способы регулирования рисков внедрении ИИ согласно стандарту ISO / IEC 23894

Риск ИИ	**Способы регулирования**
Предвзятость алгоритмов	– Оценка данных компании, на которых обучается их ИИ, на предмет исторических искажений – Использование разнообразных наборов данных – Регулярное тестирование модели на достоверность для различных демографических групп
Проблема «черного ящика»	– Использование *объяснимого ИИ* (XAI, explainable AI – системы, способные о объяснять свои действия и принимать решения понятным для людей образом [15]) – Использование *локальных интерпретируемых объяснений*, не зависящих от модели (LIME, Local Interpretable Model-agnostic Explanations) – Использование *аддитивных объяснений Шепли* (SHAP, SHapley Additive exPlanations – метод интерпретации машинного обучения, который объясняет вклад каждого признака в предсказание конкретного наблюдения [13])
Конфиденциальность данных	– Применение методов повышения конфиденциальности: *дифференциальная приватность* (DP, differential privacy – метод, обеспечивающий точные запросы в статистическую базу данных при одновременной минимизации идентификации записей в ней [6]), *федеративное обучение* (federated learning – заключение модели в защищенную среду и ее обучение без перемещения данных куда-либо [14]) и *гомоморфное шифрование* (FHE – Fully Homomorphic Encryption, вычисление на зашифрованных данных без их дешифровки [11]) – Применение принципа минимизации данных (сбор и хранение только тех данных, которые необходимы для достижения целей использования ИИ)

Продолжение Таблицы 3

Автономность и взаимодействие с физическим миром	– Использование формальной верификации (formal verification – доказательство с использованием математических методов корректности программного обеспечения [9]) – Использование сценарного тестирования (scenario-based testing) – Информирование о том, что программа или устройство с ИИ имеет системные ограничения и требует надзора человека)

Законодательные инициативы в области ИИ активно обсуждаются на уровне правительств во многих странах. Так, Европейский парламент и Европейский совет согласовали положения Закона об искусственном интеллекте (AI Act), в основе которого – риск-ориентированный подход. Из основных положений можно выделить следующие: градация систем ИИ по уровню риска, запрет вредоносных систем ИИ, регистрация высокорисковых систем ИИ в европейской базе данных до ввода в эксплуатацию, создание Управления по ИИ для развития стандартов и тестирования, установление штрафов за нарушение правил, создание «регуляторной песочницы».

В то же время в США 30 октября 2023 г. принят Указ о безопасном, надежном и заслуживающем доверия ИИ со следующими основными положениями: необходимость сообщать правительству США информацию о результатах тестирования безопасности систем ИИ, разработка стандартов, повышение безопасности персональных данных, запуск Национального ресурса исследований в области ИИ, усиление международного сотрудничества.

В Китае 10 июля 2023 г. приняты Временные меры по управлению генеративными системами ИИ. Основными положениями документа являются: соответствие систем ИИ социалистическим ценностям, ответственность разработчиков за генерируемый контент, сотрудничество между различными организациями в сфере ИИ, укрепление международного сотрудничества, создание механизма подачи обратной связи.

В России рекомендации по управлению рисками ИИ обозначены в Кодексе этики искусственного интеллекта [3]. Так, компаниям, использующим системы ИИ, рекомендовано проводить оценку потенциальных рисков применения ИИ, в том числе – с помощью независимого аудита, и выработать соответствующие методики оценки рисков. Кроме того, в документе отмечено, что решения в области применения ИИ должно сопровождаться научно выверенным, междисциплинарным прогнозированием социально-экономических последствий и рисков, изучением возможных изменений в ценностно-культурной парадигме развития общества с учетом национальных приоритетов.

Помимо кодекса в России приняты законы, связанные с регулированием использования ИИ-технологий. Так, 9 июля 2024 г. в России введена ответственность за причинение вреда при использовании решений с ИИ. Этот шаг стал результатом внесения поправок в Федеральный закон «Об экспериментальных правовых режимах в сфере цифровых инноваций в Российской Федерации» [6]. По новым положениям закона предусмотрено страхование рисков, возникающих при использовании технологий ИИ, что обеспечивает дополнительную защиту для граждан и юридических лиц.

Также 26 июля 2024 г. Совет по правам человека при президенте РФ и Минцифры России сообщили о совместной инициативе по разработке правил и ограничений использования ИИ в здравоохранении, образовании, судопроизводстве, транспорте, сфере безопасности и психологической помощи [4].

Одним из примеров работы с общественным сознанием является американская программа по развитию цифровой грамотности Digital Equity. В рамках программы создаются курсы по ИИ-грамотности, лаборатории и онлайн-платформы, которые облегчают обучение в области ИИ. Программа развивается в рамках законопроекта AI Literacy Act, который поддерживает широкий круг образовательных ассоциаций и организаций, например

Американская федерация учителей, ETC, Intel, объединение TeachAI, Университет штата Делавэр и другие.

В России программы цифровой грамотности реализуются и государством, и частными организациями. Так, в январе 2024 г. «Яндекс Практикум» запустил бесплатную программу «Цифровая грамотность и безопасность в интернете», в которой можно узнать больше об ИИ, безопасном поиске, работе с онлайн-сервисами, использовании нейросетей и облачных хранилищ.

Таким образом, внедрение ИИ существенно трансформирует различные аспекты общества, создавая серьезные риски. Негативные последствия внедрения ИИ могут проявляться в нескольких сферах: от экономических проблем, связанных с автоматизацией рабочих мест, до социальных и этических вызовов, таких как нарушение прав человека и углубление неравенства. Понимание и управление этими рисками требует междисциплинарного подхода, сочетающего научные исследования, этическую оценку и политическое регулирование.

Экономические последствия, такие как сокращение рабочих мест в ряде отраслей и рост неравенства, требуют активного вмешательства государства, включая разработку программ переподготовки и поддержки работников. Этические и социальные последствия требуют создания четких нормативных стандартов, которые бы регулировали использование ИИ и обеспечивали защиту прав человека и социальных норм. Важную роль в этом процессе играет просвещение общества и повышение уровня цифровой грамотности, что способствует более осознанному и ответственному использованию технологий.

Кроме того, эффективное управление социальными рисками требует международного сотрудничества и разработки глобальных стандартов в области ИИ. Координация усилий на международном уровне позволит создать более согласованную и эффективную систему регулирования, которая будет учитывать интересы всех сторон и способствовать справедливому распределению выгод и рисков, связанных с ИИ.

Таким образом, успешное управление последствиями внедрения ИИ должно опираться на комплексный подход, включающий развитие образовательных инициатив, законодательное регулирование и международное сотрудничество. Только в этом случае можно будет минимизировать негативные последствия и максимально использовать потенциал ИИ для блага общества.

Литература

1. Бразевич С.С. Влияние процессов цифровизации на систему управления социальной безопасностью российского социума / С.С. Бразевич, Я.А. Маргулян, Д.А. Багдасарян // Теория и практика общественного развития. – 2023. – № 11(187). – С. 25-32.

2. Григорьева П. А., Гаев Л. В. Культурное влияние искусственного интеллекта // Международный журнал гуманитарных и естественных наук. – 2024. – №. 6-4 (93). – С. 41-43.

3. Кодекс этики в сфере ИИ. Альянс в сфере искусственного интеллекта, 2024 [Электронный ресурс]. – URL: https://ethics.a-ai.ru (дата обращения: 05.08.2024).

4. Содействие занятости. Национальные проекты России, 2024 [Электронный ресурс]. – URL: https://xn--80aapampemcchfmo7a3c9ehj.xn--p1ai/projects/demografiya/sodeystvie_zanyatosti/ (дата обращения: 05.08.2024).

5. СПЧ и Минцифры договорились выработать правила использования ИИ. РИА Новости, 2024 [Электронный ресурс]. – URL: https://ria.ru/20240726/tekhnologii-1962282139.html (дата обращения: 05.08.2024).

6. Федеральный закон от 8 июля 2024 г. N 169-ФЗ "О внесении изменений в Федеральный закон «Об экспериментальных правовых режимах в сфере цифровых инноваций в Российской Федерации». Российская газета, 2024 [Электронный ресурс]. – URL: https://rg.ru/documents/2024/07/12/o-vnesenii-izmeneniy-v-fedaralniy-zakon-dok.html (дата обращения: 05.08.2024).

7. Что такое дифференциальная приватность? Хабр, 2016 [Электронный ресурс]. – URL: https://habr.com/ru/articles/395313 (дата обращения: 05.08.2024).

8. Шваб К. Четвертая промышленная революция / К. Шваб // Москва, «Эксмо», 2016. – 138 с.

9. A Gentle Introduction to Formal Verification. Systemverilog, 2024 [Электронный ресурс]. – URL: https://www.systemverilog.io/verification/gentle-introduction-to-formal-verification/ (дата обращения: 05.08.2024).

10. Beyond ISO 42001: The Role of ISO/IEC 23894 in AI Risk Management. Medium, 2024 [Электронный ресурс]. – URL: https://medium.com/@mukherjee.amitav/beyond-iso-42001-the-role-of-iso-iec-23894-in-ai-risk-management-7c4f3036544f (дата обращения: 05.08.2024).
11. Federated Learning meets Homomorphic Encryption. IBM, 2024 [Электронный ресурс]. – URL: https://research.ibm.com/blog/federated-learning-homomorphic-encryption (дата обращения: 05.08.2024).
12. Levy R. Social media, news consumption, and polarization: Evidence from a field experiment //American economic review. – 2021. – Т. 111. – №. 3. – С. 831-870.
13. Ribeiro M. T. Why should I trust you? Explaining the predictions of any classifier / Ribeiro M. T., Singh S., Guestrin C. // Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. – 2016. – С. 1135-1144.
14. Trask A. W. Grokking deep learning. Simon and Schuster, 2019.
15. What is explainable AI? IBM, 2024 [Электронный ресурс]. – URL: https://www.ibm.com/topics/explainable-ai (дата обращения: 05.08.2024).

Школа молодых исследователей

DOI: 10.24412/1994-3776-2024-3-186-194 УДК: 332.14

А.О. Федоровский

Реализация проекта Smart City Berlin как пример эффективного управления инновационно-привлекательным городом

A. Fedorovskiy. The implementation of the Smart City Berlin project as an example of effective management of an innovative and attractive city

Аннотация. Статья рассматривает проект Smart City Berlin как пример успешного управления городом с помощью инновационных технологий. Берлин активно внедряет цифровые решения, улучшая качество жизни, устойчивое развитие и решая городские проблемы. Примеры инициатив показывают, как город справляется с вызовами в сферах транспорта, экологии и социальных услуг. Важной особенностью является система, активно включающая граждан в управление городом. Успех Берлина подтверждается высокими позициями в мировых рейтингах, подчеркивая его роль как ведущего инновационного центра.

Ключевые слова: Умный город, Берлин, Германия, эффективное управление, цифровизации, устойчивое развитие, привлекательность города, технологии управления.

Контактные данные: 117624, г. Москва, ул. Изюмская, д. 26, +79853197640, e-mail: arseniy.fedorovskiy@mail.ru

Abstract. The article considers the Smart City Berlin project as an example of successful city management using innovative technologies. Berlin is actively implementing digital solutions, improving the quality of life, sustainable development and solving urban problems. Examples of initiatives show how the city copes with challenges in the fields of transport, ecology and social services. An important feature is the system that actively involves citizens in the management of the city. Berlin's success is confirmed by its high positions in world rankings, emphasizing its role as a leading innovation center.

Keywords: Smart city, Berlin, Germany, effective management, digitalization, sustainable development, attractiveness of the city, management technologies.

Contact information: 26 Izumskaya str., Moscow, 117624, +79853197640, e-mail: arseniy.fedorovskiy@mail.ru

Федоровский Арсений Олегович – магистрант Национального исследовательского университета «Высшая Школа Экономики»
A. Fedorovskiy– master's student of Higher School of Economics