Вестник МИРБИС. 2025. № 2 (42)′. С. 53–64.
Vestnik MIRBIS. 2025; 2 (42)′: 53–64.

Научная статья
УДК: 004.8:004.056.53:34.096
DOI: 10.25634/MIRBIS.2025.2.7

Управление рисками продуктов, основанных на генеративном искусственном интеллекте

Анна Александровна Островская[1], **София Денисовна Денисова**[1,2], **Виктор Евгеньевич Машинистов**[1], **Александр Юрьевич Оленев**[1]

Аннотация. Актуальность исследования обусловлена стремительным развитием генеративного искусственного интеллекта (ИИ) в России и отсутствием устоявшихся практик управления связанными с ним рисками на уровне продуктовых команд. Несмотря на рост рынка ИИ-решений и интерес со стороны бизнеса, нормативная база остается недостаточно разработанной, что повышает ответственность разработчиков и менеджеров продуктов за минимизацию потенциальных угроз. Цель статьи — исследование ключевых рисков продуктов на базе генеративного ИИ и подходов к управлению ими в российском контексте. В качестве материалов использованы научные и деловые публикации, нормативные документы, а также данные экспертной оценки рисков методом Дельфи, апробированным при разработке ИИ-продукта в сфере психологической поддержки. Для иллюстрации рисков и управления ими приведены примеры на основе бенчмарк-анализа чат-ботов психологической поддержки. Ведущим подходом к исследованию является анализ и синтез информации, дополненный методом экспертной оценки, что позволило комплексно рассмотреть технологические, этические, юридические и пользовательские риски. В статье представлены пять ключевых рисков: конфиденциальность и безопасность данных, этические проблемы, качество ответов ИИ, юридическая ответственность и отсутствие объяснимости. Раскрыты меры по управлению каждым из них, включая технические и организационные решения. Результаты исследования подтвердили гипотезу о недостатке практико-ориентированных руководств по управлению рисками генеративного ИИ для российского рынка. Материалы статьи представляют практическую ценность для менеджеров продуктов и разработчиков, предлагая структурированные подходы к минимизации рисков и повышению надежности ИИ-решений.

Ключевые слова: искусственный интеллект, управление продуктом, риск-менеджмент, этика, конфиденциальность.

Для цитирования: Управление рисками продуктов, основанных на генеративном искусственном интеллекте / А. А. Островская, С. Д. Денисова, В. Е. Машинистов, А. Ю. Оленев. DOI: 10.25634/MIRBIS.2025.2.7 // Вестник МИРБИС. 2025; 2:53–64.

JEL: M15, M40, O33

Original article

Risk Management for GenAI-Based Products

Anna A. Ostrovskaya[3], **S. D. Denisova**[3,4], **Viktor E. Mashinistov**[3], **Aleksandr Yu. Olenev**[3]

Abstract. The relevance of the study is driven by the rapid development of generative artificial intelligence (AI) in Russia and the lack of established practices for managing associated risks at the product team level. Despite the growth of the AI solutions market and business interest, the regulatory framework remains underdeveloped, increasing the responsibility of developers and product managers to mitigate potential threats. The aim of the article is to explore the key risks of products based on generative AI and approaches to managing them in the Russian context. The materials used include scientific and business publications, regulatory documents, and data from a risk assessment conducted using the Delphi method, tested during the development of an AI-based psychological support product. Examples from a benchmark

1 Российский университет дружбы народов имени Патриса Лумумбы, Москва, Россия.
2 sofden1@yandex.ru
3 Peoples' Friendship University of Russia named after Patrice Lumumba, Moscow, Russia.
4 sofden1@yandex.ru

analysis of psychological support chatbots are provided to illustrate the risks and their management. The leading research approach involves the analysis and synthesis of information, supplemented by expert evaluation, enabling a comprehensive examination of technological, ethical, legal, and user-related risks. The article presents five key risks: data confidentiality and security, ethical issues, the quality of AI responses, legal liability, and lack of transparency. Measures to manage each of these risks are outlined, including technical and organizational solutions.
The study's results confirm the hypothesis about the lack of practical guidelines for managing generative AI risks in the Russian market. The article's findings hold practical value for product managers and developers, offering structured approaches to minimizing risks and enhancing the reliability of AI solutions.

Key words: artificial intelligence, product management, risk management, ethics, confidentialityn.

For citation: Risk Management for GenAI-Based Products. By Anna A. Ostrovskaya, S. D. Denisova, V. E. Mashinistov, A. Yu. Olenev. DOI: 10.25634/MIRBIS.2025.2.7. *Vestnik MIRBIS*. 2025; 2:53–64 (in Russ.).

JEL: M15, M40, O33

Введение

В последние годы в России наблюдается стремительный рост интереса к искусственному интеллекту со стороны государства и бизнеса.

При этом помимо цифровой трансформации компаний и отраслей, важным аспектом масштабного развития ИИ-технологий в России является появление и рост отдельного рынка ИИ-продуктов, преимущественно в B2B-сегменте. Совокупная выручка компаний от продажи ИИ-решений для бизнеса в 2022 году оценивается в 30–50 млрд руб. в год, а к 2028 году в позитивном сценарии эта цифра может вырасти до 0,3–0,6 трлн руб[2]. Кроме того, согласно опросам, 36 % продакт-менеджеров хотели бы создавать продукты на базе ИИ[3], что доказывает привлекательность данного направления для профессионалов. В данном направлении особую роль играет генеративный ИИ, который уже применяется, согласно опросу VK и Prognosis, 70 % участниками отечественного рынка[4].

По мере повышения зрелости технологий, роль искусственного интеллекта смещается от поддержки принятия решений к автономности, когда ответственность за создание ценности делегируется ИИ-системам. Этот тренд несёт с собой не только преимущества в виде скорости и масштабируемости, но и порождает качественно новые типы рисков: технологических, этических, правовых, пользовательских и других.

При этом, несмотря на наличие рекомендательных документов (таких как Кодекс этики в сфере ИИ и иных документов Альянса в сфере ИИ), в России отсутствует нормативная база (за исключением законопроекта, который находится на стадии формирования и обсуждения[5]) и устоявшиеся практики управления этими рисками в рамках продуктовых команд.

Таким образом, несмотря на то что на текущий момент управление рисками остается на усмотрение разработчиков и провайдеров продуктов на базе ИИ, существует недостаточно информации о том, как следует рассматривать данные риски в призме продуктового менеджмента и разрабатывать, внедрять и регулярно совершенствовать меры по управлению ими на уровне продукта.

Всё это подтверждает высокую актуальность темы работы и перспективность данного направления исследования.

Данная статья вносит вклад в преодоление данного разрыва и нацелена на исследование ключевых рисков продуктов на базе генератив-

---

1 © А. А. Островская, С. Д. Денисова, В. Е. Машинистов, А. Ю. Оленев, 2025
Вестник МИРБИС, 2025, № 2 (42), с. 19–26.

2 Искусственный интеллект в России – 2023: тренды и перспективы / Яков и Партнеры, Яндекс. Москва, 2023. 80 с. Текст : электронный. URL: https://yakov.partners/upload/iblock/c5e/c8t1wrkdne5y9a4nqlicderalwny7xh4/20231218_AI_future.pdf (дата обращения: 01.05.2025).

3 Треть продакт-менеджеров России хочет создавать продукты на базе ИИ. Текст : электронный // CNews : новостной сайт. URL: https://www.cnews.ru/news/line/2025-05-22_tret_prodakt-menedzherov (дата обращения: 01.05.2025).

4 Генеративный искусственный интеллект используют 70 % российских компаний. Текст : электронный // Forbes : сайт. URL: https://www.forbes.ru/tekhnologii/535047-generativnyj-iskusstvennyj-intellekt-ispol-zuut-70-rossijskih-kompanij (дата обращения: 01.05.2025).

5 В России обсудят запрет ИИ с «угрожающим уровнем риска». Кто должен нести ответственность за вред, нанесенный технологией человеку. Текст : электронный // РБК : официальный сайт консалтинговой компании. URL: https://www.rbc.ru/technology_and_media/11/04/2025/67f7dc399a79477fdd97bf30 (дата обращения: 01.05.2025).

ного ИИ и подходов к управлению рисками, связанными с искусственным интеллектом.

Для достижения поставленной цели были сформулированы и выполнены следующие задачи:

1. Описать ключевые риски, связанные с применением генеративного ИИ в основе продукта, актуальные в российском контексте, включая возможные меры управления ими.
2. Проанализировать существующие подходы к управлению рисками искусственного интеллекта с точки зрения их применимости в рамках управления продуктом, в том числе на примере оценки рисков для ИИ-продукта в сфере психологи.
3. Проанализировать примеры практик по управлению рисками генеративного ИИ в отечественных цифровых продуктах на примере чат-ботов психологической поддержки.

Выбор продуктов в области психологии для иллюстрации теоретических положений работы обусловлен опытом авторов в разработке аналогичного решения, а также высокой значимостью рисков ИИ для таких продуктов, бизнес-модель которых полностью зависит от больших языковых моделей, обеспечивающих работу чат-бота.

Гипотеза, поставленная в начале исследования, заключается в том, что, несмотря на наличие различных регуляторных инициатив в области ИИ за рубежом и в России, а также имеющихся научных и деловых источников на данную тему, на текущий момент не существует в открытом доступе ориентированного на практику и адаптированного для российского контекста руководства по управлению рисками продукта, основанного на генеративном ИИ.

Практическая ценность статьи заключается в агрегированном представлении информации о рисках, связанных с построением продукта на базе ИИ-технологий, и подходах к управлению ими, что может использоваться менеджерами продукта и командой разработки продукта.

Методологические основы

В статье используются методы анализа и синтеза информации из научной и деловой литературы, а также метод экспертной оценки рисков (метод Дельфи), апробированный при разработке ИИ-продукта в сфере психологической поддержки «AI Psychology».

В качестве наглядных примеров того, каким образом продукт внедряет или не внедряет лучшие практики по управлению рисками ИИ, приведены сформированные авторами статьи наблюдения по результатам проведенного бенчмарк-анализа продуктов-аналогов для B2C-сегмента, функционально зависимых от генеративного ИИ (чаще всего — от LLM ChatGPT от OpenAI): Виртуальный психолог Лея (https://leahelps.ru/), ИИ-психолог «Мира» (https://mirahelps.online/), онлайн-психолог Анна (https://t.me/calm_mind_bot), бот-психолог Ася (https://getasya.ru/).

Результаты

1. Обзор ключевых рисков генеративного ИИ на уровне продукта

На основе проведенного исследования и анализа существующей научной и деловой литературы, обзоров и докладов авторитетных организаций в области цифровой трансформации, были выявлены пять ключевых рисков, связанных с применением генеративного искусственного интеллекта в продукте. Рассмотрим детально каждый из этих рисков с указанием описания риска, его последствий, а также возможных мер по управлению им.

1.1. Риски конфиденциальности и безопасности данных

Ключевой вызов любого цифрового продукта, в том числе продуктов на основе искусственного интеллекта, — это защищенность персональных данных пользователей [Jobin, 2019].

Утечка, взлом или неправомерное использование данных пользователей может привести к юридическим последствиям (штрафы, судебные иски за нарушение 152-ФЗ[1] и других регуляторных норм[2]), репутационным потерям (подрыв доверия пользователей и заказчиков, негативные публикации в СМИ), финансовым убыткам (компенсации пострадавшим, затраты на устранение последствий утечки). Особенно уязвимы продукты, где пользователи вводят чувствительные данные, в том числе медицинские — как, например, в сервисах заботы о здоровье.

1 Российская Федерация : Федеральный закон : О персональных данных от 27.07.2006 г. №152-ФЗ, в редакции Федерального закона Российской Федерации от 08.08.2024 № 233-ФЗ // Собрание законодательства РФ, 31.07.2006, №31 (1 ч), ст. 3451.

2 Персональные данные: новые штрафы с 30 мая 2025 года. Текст : электронный // КонсультантПлюс. URL: https://www.consultant.ru/legalnews/28492/ (дата обращения: 01.05.2025).

Известным примером реализации риска является произошедшая весной 2023 года утечка данных корпорации Samsung, содержащих корпоративную коммерческую тайну, из-за применения сотрудниками в работе ChatGPT [Былевский, 2023]. Поскольку OpenAI использует введённые данные для обучения своих моделей, то сведения об исходном коде новых программ, повестке внутренних совещаний и информации о новом оборудовании могли быть сохранены и потенциально использованы. После выявления инцидента Samsung ограничила использование ChatGPT и начала разработку собственной ИИ-системы.

В общем виде на уровне продукта меры по управлению данным риском включают как технические, так и организационные мероприятия. Среди технических мер: сквозное шифрование данных (при их передаче и при хранении), изоляция баз данных с персональной информацией от других корпоративных систем, аутентификация по SSO (Single Sign-On), регулярные независимые аудиты безопасности. К организационным мерам можно отнести следующие: четкие политики доступа (принцип минимальных привилегий), проверка сотрудников на знание основ информационной безопасности, юридическое сопровождение (соответствие 152-ФЗ), публичная политика конфиденциальности и заявления о безопасности во всех каналах коммуникации, рекламных сообщениях и пользовательских соглашениях сервиса, анонимизация данных везде, где это возможно (например, для аналитики).

Характеризуя различный уровень соответствия требованиям конфиденциальности, можно привести в пример политики конфиденциальности двух сервисов психологической поддержки на основе чат-ботов. «Политика обработки персональных данных и конфиденциальности в отношении обработки персональных данных» сервиса «Мира» содержит детальное описание различных аспектов работы с персональными данными пользователей (ПДн), в том числе адрес электронной почты для направления заявлений в отношении персональных данных пользователей, особенности хранения, блокировки, удаления и иных операций с ПДн[1]. С другой стороны, в рамках продукта «Бот Лея» на официальном сайте представлена ссылка «Политика конфиденциальности регулируется мессенджером Telegram: https://telegram.org/privacy-tpa» без раскрытия обязательных согласно 152-ФЗ аспектов[2].

1.2. Этические риски

ИИ-модель может воспроизводить или усиливать социальные, гендерные, культурные стереотипы; некорректные рекомендации; дискриминационные паттерны [Волобуев, 2023].

Результатом являются репутационные потери — обвинения в неэтичности, публичные скандалы, потеря доверия клиентов; юридические следствия (возможные иски за дискриминацию), а также негативные финансовые эффекты в виде оттока клиентов, затрат на доработку и т.п.

Для управления этическими рисками предусмотрен широкий набор возможных мер. На уровне контроля данных для обучения моделей машинного обучения рекомендуется использовать сбалансированные датасеты (разные демографические группы, культурные контексты) и фильтровать токсичный и стереотипный контент на этапе подготовки данных. После обучения и в процессе эксплуатации моделей в рамках модерации их работы предлагается проводить аудит ответов бота (регулярные проверки независимыми тестировщиками), донастраивать модели для соответствия нормам.

Вопросы аудита систем и моделей ИИ, в особенности больших языковых моделей, имеют значительную глубину и методологическую проработку в научной литературе. В частности, одно из предложений исследователей предполагает трехуровневый подход к аудиту крупных языковых моделей, включая аудит провайдеров технологий, самих моделей и их приложения [Mökander, 2023]. Также отдельная дискуссия ведется относительно выбора между внутренним и внешним аудита, преимуществ и недостатков каждого из вариантов [Krause, 2025].

Техническими решениями для управления этическими рисками являются детекторы (алгоритмы) стереотипных суждений ИИ, а также жесткие запреты на предоставление рекомендаций по запрещенным темам, таким как медицина, ре-

1 Политика обработки персональных данных и конфиденциальности в отношении обработки персональных данных. Текст : электронный // Мира : сайт. URL: https://mirahelps.online/privacyPolicy.pdf (дата обращения: 01.05.2025).

2 Политика конфиденциальности. Текст : электронный // Бот Лея : сайт. URL: https://leahelps.ru/privacy-policy/ (дата обращения: 01.05.2025).

лигия, политика и другие. Важно предоставлять прозрачную информацию об использовании ИИ и его функционале в продукте пользователям, раскрывая ограничения ИИ, а также создавая механизм жалоб на некорректные ответы ИИ.

1.3. Качество и достоверность ответов ИИ

Одним из ключевых рисков, с которым сталкиваются разработчики ИИ-продуктов, является снижение качества и достоверности выдаваемых моделью рекомендаций. Наиболее уязвимыми аспектами качества являются некорректные или вредные советы, выдаваемые в ответ на запросы в критических ситуациях.

Дополнительное усложнение возникает из-за потери контекста в диалоге. Архитектуры типа GPT обладают ограниченным «окном внимания» — числом токенов, которые модель может учитывать при генерации ответа. При увеличении длины диалога важные детали могут быть вытеснены или забыты, что приводит к потере персонализации, фрагментарности диалога и появлению ощущений повторяемости и оторванности от реального запроса. Это особенно чувствительно в ситуациях, требующих накопления и анализа индивидуального опыта пользователя. Также проблема качества проявляется в чрезмерной многословности и слабой структурированности ответов. Генеративные модели нередко создают перегруженные текстовые блоки, содержащие множество тезисов или рекомендаций одновременно, что затрудняет их восприятие и снижает эффективность коммуникации.

Управление данным риском требует системных мер, включающих техническую оптимизацию диалоговой архитектуры, повышение контекстной устойчивости модели, разработку механизмов персонализации и уточнения выводов на каждом этапе общения. Существенным элементом являются технологии препроцессинга и постпроцессинга, обеспечивающие сжатие, структурирование и логическое оформление ответов модели. На уровне инфраструктуры эффективными практиками являются хранение ключевых пользовательских параметров вне основного окна внимания модели (в рамках защищённой внешней памяти) с соблюдением требований законодательства о персональных данных. В дополнение к этому, важна регулярная калибровка модели доменными экспертами, а также системный сбор и анализ обратной связи от пользователей, позволяющий выявлять слабые места в коммуникации. Отдельное внимание следует уделить пользовательскому опыту.

1.4. Юридическая ответственность

Риск юридической ответственности представляет собой одну из наиболее значимых и комплексных угроз для разработчиков и операторов универсальных продуктов на базе генеративного искусственного интеллекта, особенно в случаях, когда ИИ-система вступает в диалог с пользователями по вопросам, затрагивающим здоровье, благополучие или иные чувствительные сферы.

Прежде всего, актуальна проблема некорректной квалификации сервиса: несмотря на то, что продукт может позиционироваться как информационный или поддерживающий, при определённых сценариях использования он может быть интерпретирован как средство оказания медицинских или психологических услуг. Такая ситуация влечёт за собой высокие регуляторные риски, включая необходимость лицензирования, соблюдения профессиональных стандартов и привлечения сертифицированных специалистов. В случае отсутствия соответствующего правового оформления, компания может быть признана нарушителем законодательства, в частности, в рамках статьи 238 УК РФ, предусматривающей ответственность за оказание услуг, не отвечающих требованиям безопасности.

Дополнительным аспектом риска является потенциальная ответственность за вред, причинённый пользователю вследствие следования советам, сгенерированным ИИ. Если в результате таких рекомендаций произойдёт усугубление состояния человека (например, в контексте депрессивных эпизодов, тревожных расстройств или других нарушений), юридические последствия могут включать гражданские иски, а также административные или уголовные санкции. Особенно чувствительными в этом отношении являются случаи, в которых пользователь может трактовать взаимодействие с ботом как замену консультации с квалифицированным специалистом, что особенно вероятно при отсутствии четких разграничений статуса сервиса.

С целью минимизации данных рисков критически важным является обеспечение прозрачного юридического позиционирования продукта. Это включает наличие явных уведомлений о том, что ИИ-система не является медицинским или психи-

атрическим инструментом, не предназначена для диагностики или лечения, и не заменяет профессиональную помощь. Дополнительно требуется разработка механизмов автоматического реагирования на критические запросы, включая перенаправление пользователя к специализированным службам или предложению срочной помощи. Юридическая и психологическая экспертиза всех скриптов и формулировок, используемых в продукте, должна быть обязательной частью процесса валидации перед запуском, как и наличие договорной базы с корпоративными клиентами, чётко разграничивающей зоны ответственности сторон. Для повышения устойчивости к возможным претензиям также рекомендуется страхование профессиональной ответственности, позволяющее покрыть потенциальные судебные издержки.

Например, один из наиболее популярных бесплатных чат-ботов психологической поддержки в Telegram «Онлайн-психолог Анна» имеет ежемесячную аудиторию свыше 11 тыс. человек, при этом у данного продукта отсутствует сайт, не представлена какая-либо формализованная документация о продукте, включая политику конфиденциальности и иные юридические документы[1]. Вся доступная информация о продукте размещается в специальном канале в Telegram, аудитория которого превышает 1,5 тыс. человек[2]. Таким образом, по мере роста аудитории можно предполагать и о усилении юридических рисков данного продукта, предназначенного для психологической поддержки.

1.5. Отсутствие объяснимости и прозрачности ИИ

Риск отсутствия объяснимости и прозрачности генеративных моделей ИИ является фундаментальной проблемой. В контексте генеративного ИИ, основанного на трансформерной архитектуре, характерно отсутствие интерпретируемых механизмов принятия решений: пользователи и даже разработчики не могут с полной достоверностью понять, на каких основаниях модель пришла к тому или иному выводу. Это создает ситуацию «чёрного ящика», в которой генерируемые рекомендации могут восприниматься как произвольные или неадекватные, особенно в случаях, когда ответы не сопровождаются указанием источников, логики рассуждения или ссылками на проверенные знания [Киселева, 2024].

Отсутствие объяснимости также усугубляется динамической природой генеративных моделей, способных производить разнородные и непредсказуемые ответы в зависимости от формулировки запроса [Аверкин, 2024], что затрудняет внутреннюю проверку качества, а также обучение пользователей эффективному взаимодействию с системой. При этом непрозрачность алгоритма лишает продукт не только устойчивости к ошибкам, но и юридической и этической защищённости: невозможно ретроспективно реконструировать цепочку генерации, если потребуется обосновать корректность ответа в случае претензий. Пользователь, не получая внятных объяснений, может потерять доверие к системе, воспринимая её как непоследовательную, а рекомендации — как неавторитетные или даже потенциально вредоносные.

Последствия реализации данного риска включают снижение удовлетворённости пользователей, рост числа отказов от использования продукта и угрозу репутации, особенно если публично станет известно о непредсказуемом или неэтичном ответе, сгенерированном ИИ.

Для управления этим риском необходим комплекс технических и организационных мер. Технически возможно внедрение механизмов постобъяснения — генерация модели сопровождающего комментария, в котором указывается логика, принятая моделью (например, обоснование выбора метода психологической поддержки или отсылки к источникам знаний). Также важно разрабатывать системы, способные визуализировать внутренние процессы модели в упрощённой форме — как для разработчиков, так и для конечного пользователя. На организационном уровне следует внедрять регулярный аудит ответов с точки зрения объяснимости и логической обоснованности, а также стандарты проектирования промптов и диалоговых сценариев, обеспечивающих предсказуемость поведения системы.

Например, среди российских ИИ-решений, нацеленных на психологическую поддержку пользователей, обнаружено только одно, для которо-

1 Бот «Онлайн-психолог Анна». Текст, изображения : электронные // Telegram : социальная сеть. URL: https://t.me/calm_mind_bot (дата обращения: 01.05.2025).

2 Там же: Официальный канал бота «Онлайн-психолог Анна». URL: https://t.me/calm_mind_channel (дата обращения: 01.05.2025).

го детально раскрывается техническая архитектура и логика работы (см. рисунок).

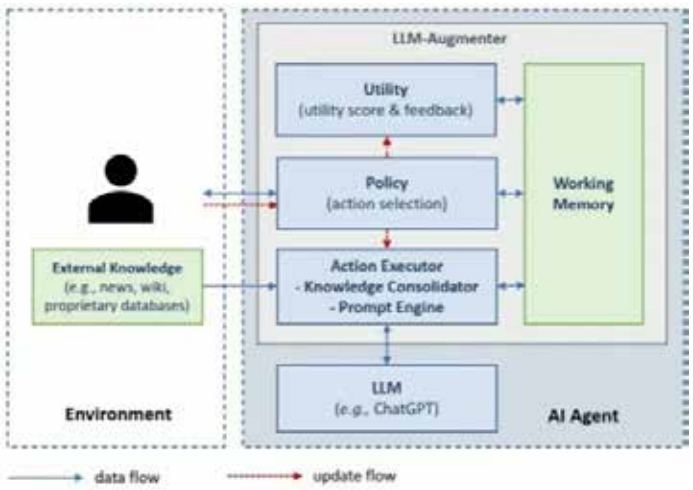

Рис. Архитектура продукта «Бот-психолог Ася»
Источник: Как работает бот-психолог «Ася» // GetAsya. ru : сайт. URL: https://getasya.ru/kak-rabotaet-bot-psiholog/ (дата обращения: 01.05.2025).

1. Анализ существующих подходов к управлению рисками ИИ

В этом разделе представлен обзор актуальных подходов к управлению рисками ИИ с акцентом на их применимость в рамках управления цифровыми продуктами. Важно указать, что работа по регулированию ИИ и минимизации рисков, связанных с ИИ в целом и генеративным ИИ в частности, ведется на различных уровнях, начиная с международного сообщества.

На глобальном уровне вопросы рисков, связанных с ИИ, стали предметом обсуждения в рамках Организации Объединённых Наций. В 2024 году Генеральная Ассамблея приняла резолюцию, призывающую к созданию глобальных стандартов безопасности и ответственности в разработке и применении ИИ-систем[1].

К числу влиятельных международных источников относится также пакет руководящих принципов, подготовленный Организацией экономического сотрудничества и развития (ОЭСР). Принципы ИИ ОЭСР (OECD AI Principles) подчеркивают необходимость подотчетности, устойчивости, транспарентности и надёжности ИИ-систем, а также предоставляют методические рекомендации по выстраиванию ответственных практик разработки и внедрения ИИ[2].

Многие государства разрабатывают собственные правовые и институциональные механизмы управления рисками ИИ. В Китае в 2024 году были утверждены Основные требования безопасности для услуг генеративного искусственного интеллекта, которые, как и подход к регулированию ИИ в Китае в целом, учитывает нормы нравственности и социалистической морали [Харитонова, 2025].

В США подходы кодифицируются в рамках Национального института стандартов и технологий (NIST), где сформирован фреймворк управления рисками ИИ (AI Risk Management Framework, NIST AI RMF)[3]. Документ создан для организаций, которые проектируют, разрабатывают, внедряют или используют системы искусственного интеллекта, с целью поддержки управления рисками, связанными с ИИ, и содействия надежной и ответственной разработке и использованию систем искусственного интеллекта. Фреймворк вводит следующие характеристики доверенного ИИ: валидный и надежный, безопасный и отказоустойчивый, конфиденциальный, прозрачный и подотчетный, интерпретируемый и объяснимый, справедливый с нулевыми негативными предубеждениями [Шейкин, 2024]. Кроме того, NIST AI RMF предлагает использование 4 функций для управления рисками ИИ.

- Стратегическое управление (Govern) — создание и развитие культуры риск-менеджмента.
- Картирование (Map) — изучение контекста и определение рисков.
- Измерение (Measure) — анализ, оценка и мониторинг идентифицированных рисков.
- Оперативное управление (Manage) — выявление приоритетных рисков и действия с учетом их прогнозируемого воздействия.

В Европейском союзе продолжается внедрение Регламента об искусственном интеллекте («AI Act»), который впервые вводит классифика-

1 В Генассамблее ООН приняли резолюцию по искусственному интеллекту. Текст : электронный // ООН : официальный сайт. URL: https://news.un.org/ru/story/2024/03/1450631 (дата обращения: 01.05.2025).

2 AI Principles. Текст : электронный // OECD : официальный сайт. URL: https://www.oecd.org/en/topics/ai-principles.html (дата обращения: 01.05.2025).

3 AI Risk Management Framework. Текст : электронный // National Institute of Standards and Technology. : официальный сайт. URL: https://www.nist.gov/itl/ai-risk-management-framework (дата обращения: 01.05.2025).

цию ИИ-систем по уровням риска и определяет юридические требования к каждой категории, включая контроль над данными и обязательную документацию для высокорисковых систем[1]. В рамках «AI Act», Европейский союз предоставляет малым и средним предприятиям (МСП) особые условия для внедрения и тестирования ИИ. Во-первых, МСП получают приоритетный и бесплатный доступ к регуляторным песочницам, где можно испытывать ИИ-продукты в реальных условиях вне обычного регулирования. Во-вторых, снижаются издержки на соблюдение требований: сборы за оценку соответствия устанавливаются пропорционально размеру компании, а Европейская комиссия обязуется регулярно их пересматривать и снижать. Также предусмотрено упрощение технической документации и специальные обучающие программы, адаптированные для МСП. Кроме того, МСП смогут участвовать в разработке стандартов и получать консультации через выделенные каналы связи. Все обязательства в отношении МСП должны быть пропорциональны масштабу и характеру их деятельности, включая специальные показатели эффективности[2]. Поскольку любой стартап, создающий продукт на основе искусственного интеллекта, по масштабам бизнеса будет отнесен к МСП, данные особенности представляют интерес и релевантны для настоящего исследования.

В России вопросы управления ИИ-рисками нашли отражение в обновлённой в 2024 году Национальной стратегии развития искусственного интеллекта до 2030 года, в том числе с учётом особенностей генеративных систем[3]. Также ведётся подготовка законопроекта, который содержит положения о запрете систем с «неприемлемым уровнем риска», введении маркировки контента и ответственности разработчиков за вред, нанесенный нейросетями. Можно предположить, что в связи с наличием на рынке крупных игроков, развивающих ИИ (таких как Сбер, Яндекс, МТС, Т-Банк и другие), законопроекты будут сфокусированы на управлении рисками, связанными с ИИ-трансформацией в бизнесе и процессах крупных корпораций. Таким образом, формирующееся законодательное поле не будет в ближайшие годы предполагать конкретных механизмов и алгоритмов действий по управлению рисками ИИ для разработчиков, владельцев и менеджеров продуктов.

Международные стандарты управления рисками, применимые к ИИ, формируют основу для системной интеграции подходов к снижению рисков в рамках жизненного цикла цифрового продукта. Прежде всего, это стандарты ISO/IEC (в России — их адаптации в виде национальных стандартов). В частности, ISO 23894:2023 по управлению рисками ИИ (и предстандарт ПНСТ 776-2022 «Информационные технологии. Интеллект искусственный. Управление рисками», адаптирующий FDIS-версию ISO/IEC 23894)[4], который основан на серии стандартов ISO 31000 по риск-менеджменту в организациях[5]. Этот стандарт акцентирует важность управления рисками на уровне дизайна, разработки, валидации и эксплуатации ИИ-систем. Стандарт ISO 23894:2023 направлен на организации, которые разрабатывают и применяют методы и системы ИИ, однако не учитывает подходов продуктового менеджмента, что подтверждает недостаточную приспособленность существующей базы подходов риск-менеджмента к продукту как к объекту менеджмента.

Тем не менее, авторами настоящей статьи был использован стандарт ГОСТ Р ИСО/МЭК 31010-2011 «Менеджмент риска. Методы оценки риска» для определения метода оценки рисков разрабатываемого продукта «AI Psychology», предоставляющего психологическую поддержку сотрудни-

1 AI Act. Текст : электронный // European Commission : официальный сайт. URL: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (дата обращения: 01.05.2025)

2 Small Businesses' Guide to the AI Act. Текст : электронный // Future of Life Institute : официальный сайт. URL: https://artificialintelligenceact.eu/small-businesses-guide-to-the-ai-act/ (дата обращения: 01.05.2025).

3 Российская Федерация : Президент РФ : Указ : О развитии искусственного интеллекта в Российской Федерации (вместе с «Национальной стратегией развития искусственного интеллекта на период до 2030 года») от 10.10.2019 № 490 (ред. от 15.02.2024). Текст : электронный // Президент России : Официальный интернет-портал правовой информации. URL: http://www.kremlin.ru/acts/bank/44731 (дата обращения: 01.05.2025).

4 ПНСТ 776-2022 «Информационные технологии. Интеллект искусственный. Управление рисками». 28 с. Текст : электронный // Росстандарт : официальный сайт. URL: https://protect.gost.ru/v.aspx?control=8&baseC=6&id=235081 (дата обращения: 01.05.2025)

5 ПНСТ 843-2023 «Информационные технологии. Стратегическое управление информационными технологиями. Последствия влияния стратегического управления при использовании искусственного интеллекта организациями». Москва : Росстандарт, 2023. 36 с. Текст :

кам и командам в B2B-модели. Был выбран описанный в стандарте метод Дельфи — метод получения экспертных оценок, которые участвуют при идентификации источников и воздействий опасности, количественной оценке вероятности и последствий и общей оценке риска. Это метод обобщения мнений экспертов, который позволяет провести независимый анализ и голосование экспертов[1].

При разработке продукта был составлен длинный список рисков продукта, который затем прошел экспертную оценку со стороны участников команды продукта в следующих ролях:

- CPO (Chief Product Officer): фокус на пользователях, репутации, удержании.
- CTO (Chief Technology Officer): техническая реализуемость, безопасность, масштабируемость.
- COO (Chief Operating Officer): операционные затраты, экономика, внедрение.
- Применялись следующие критерии оценки:
- Вероятность риска (P): низкая (1), средняя (2), высокая (3).
- Влияние риска (I): низкое (1), среднее (2), высокое (3).
- Уровень риска (R) = P × I (1–9):

1–3: приемлемый риск;
4: требует контроля;
5: средний риск;
6: высокий риск;
7–9: критический риск.

Результаты представлены в Таблице.

Таблица. **Экспертная оценка рисков продукта «AI Psychology»**

Риск	CPO (P/I/R)	CTO (P/I/R)	COO (P/I/R)	Итоговый R	Приоритет
1) Конфиденциальность и безопасность	2/3/6	3/3/9	2/3/6	7	Критический
2) Этические риски ИИ	2/3/6	2/2/4	1/2/2	4	Контроль
3) Качество ответов ИИ	3/3/9	2/2/4	2/3/6	6	Высокий
4) Технологические трудности	1/2/2	3/2/6	3/3/9	6	Высокий
5) Юридическая ответственность	3/3/9	1/3/3	2/3/6	6	Высокий
6) Недоверие пользователей	3/3/9	1/2/2	2/3/6	6	Высокий
7) Оценка эффективности	2/2/4	1/1/1	3/2/6	4	Контроль
8) Усугубление проблем	3/3/9	2/2/4	1/3/3	5	Средний
9) Эффективность кастомизации	1/2/2	2/2/4	3/3/9	5	Средний

Источник: составлено авторами

По итогам упражнения выявлены риски, которые играют решающую роль (критический и высокий уровень риска):

1. Конфиденциальность и безопасность
2. Качество ответов ИИ
3. Технологические трудности
4. Юридическая ответственность
5. Недоверие пользователей

Именно на эти пункты будет сделан фокус при разработке и реализации дорожной карты мероприятий по управлению рисками.

Крупные технологические корпорации также разрабатывают собственные рамки и инструменты для минимизации рисков. Например, Google представила платформу SAIF (Secure AI Framework), обеспечивающую структурный подход к внедрению ИИ с учетом кибербезопасности, доступов, мониторинга моделей и управления инцидентами[2].

В России значимым шагом стал Кодекс этики в сфере ИИ[3], а также Декларация об ответственной разработке и использовании сервисов генера-

электронный // Росстандарт : официальный сайт. URL: https://docs.cntd.ru/document/1304145336 (дата обращения: 01.05.2025).

1 ГОСТ Р ИСО/МЭК 31010-2011. Менеджмент риска. Методы оценки риска. Москва : Стандартинформ, 2012. 50 с. Текст : электронный // Электронный фонд правовых и нормативно-технических документов : сайт. URL: https://docs.cntd.ru/document/1200090083 (дата обращения: 01.05.2025).

2 Secure AI Framework от Google (SAIF). Текст : электронный // Центр безопасности Google : сайт. URL: https://safety.google/cybersecurity-advancements/saif/ (дата обращения: 01.05.2025).

3 Кодекс этики в сфере искусственного интеллекта / Альянс в сфере искусственного интеллекта — Комиссия по реализации Кодекса этики

тивного ИИ[1], подписанная рядом ведущих технологических организаций. Эти документы фиксируют обязательства по обеспечению прозрачности, недопущению дискриминации, а также уважению частной жизни и конфиденциальности пользователей.

Обсуждение

Проблема интеграции риск-менеджмента и продакт-менеджмента, в особенности применительно к рискам искусственного интеллекта, является слабо исследованной в российском научном пространстве.

Анализируя актуальные публикации, можно обнаружить исследования, посвященные инновационным рискам, в том числе искусственного интеллекта, и рискам цифровой трансформации, преимущественно относящимся к организации (бизнесу), под которым подразумевается зрелый бизнес, а не стартап [Головина 2022; Кулакова 2022].

Новизна настоящей работы заключается в рассмотрении рисков, связанных с искусственным интеллектом, через призму подходов продуктового менеджмента. Формирование конкретных, научно обоснованных, но при этом практико-ориентированных и эмпирически подтвержденных мер по управлению рисками ИИ в парадигме управления продуктом, является крайне перспективным для российского рынка. Аналогичные работы имеют место за рубежом — например, [Lin 2025].

Заключение (Выводы)

Проведенное исследование подтвердило гипотезу о том, что в России отсутствуют адаптированные для практического применения руководства по управлению рисками продуктов на базе генеративного ИИ. Несмотря на наличие международных и российских регуляторных инициатив, их интеграция в продуктовый менеджмент остается слабо изученной. Основные результаты работы включают:

1. Выявление пяти ключевых ИИ-рисков, требующих первоочередного внимания: конфиденциальность данных, этические проблемы, качество ответов ИИ, юридическая ответственность и прозрачность ИИ.
2. Разработку мер по управлению рисками, сочетающих технические решения и организационные подходы.
3. Демонстрацию применимости метода Дельфи для оценки рисков на примере ИИ-продукта в сфере психологии.
4. Использование примеров из сравнительного анализа серии решений чат-ботов психологической поддержки для обоснования необходимости создания и распространения практических руководств по ИИ-риск-менеджменту на уровне продукта.
5. Описание существующих инициатив по регулированию ИИ-систем и оценку их применимости к российским продуктам.

Перспективы дальнейших исследований связаны с углубленным изучением практик внедрения риск-менеджмента в продуктовые команды, а также адаптацией международных стандартов (таких как NIST AI RMF) к российскому контексту. Статья вносит вклад в формирование методологической базы для управления рисками генеративного ИИ, что особенно актуально в условиях роста рынка и усиления регуляторного внимания.

Список источников

1. Аверкин 2024 — *Аверкин А. Н.* Объяснительный искусственный интеллект в больших речевых моделях. EDN: ROFNZG // Речевые технологии. 2024;1:3–13. ISSN: 2305-8129; eISSN: 2305-8137.
2. Былевский 2023 — *Былевский П. Г.* Культурологическая деконструкция социальнокультурных угроз ChatGPT информационной безопасности российских граждан. DOI: 10.7256/2454-0757.2023.8.43909. EDN: UZDRFW // Философия и культура = Philosophy and Culture. 2023; 8:46–56. ISSN: 1999-2793; eISSN: 2454-0757.
3. Волобуев 2023 — *Волобуев А. В.* Этика искусственного интеллекта, дискриминация и неравенство. DOI: 10.30884/vglob/2023.03.04. EDN: DNUTN // Век глобализации. 2023; 3: 48-62. (ISSN: 1994-9065.

в сфере искусственного интеллекта, 2025.10 с. Текст : электронный // Альянс в сфере искусственного интеллекта : официальный сайт. URL: https://ethics.a-ai.ru/assets/ethics_files/2023/05/12/%D0%9A%D0%BE%D0%B4%D0%B5%D0%BA%D1%81_%D1%8D%D1%82%D0%B8%D0%BA%D0%B8_20_10_1.pdf (дата обращения: 01.05.2025).

1 Там же: Декларация об ответственной разработке и использовании сервисов в сфере генеративного искусственного интеллекта. 4 с. URL: https://ethics.a-ai.ru/assets/ethics_files/2024/03/13/GenAi_Declaration_Ai_Alliance_Russia_FpNJ2Lc_82yB8pD.pdf (дата обращения: 01.05.2025).

4. Головина 2022 — *Головина Т. А.* Управление рисками организаций в условиях цифровой экономики / Т. А. Головина, И. Л. Авдеева, Д. А. Суханов. DOI: 10.24412/2304-6139-2022-48-1-55-61. EDN: CXBFBR // Вестник академии знаний. 2022; 48(1): 55–61. ISSN: 2304-6139; eISSN: 2687-0983.
5. Киселева 2024 — *Киселева Л. А.* Правовое регулирование искусственного интеллекта: от «черного ящика» к прозрачности алгоритмов / Л. А. Киселева, А. С. Шайдуров. EDN: LONHBK // Цивилизационные перемены в России : материалы XIV Всероссийской научно-практической конференции, Екатеринбург, 29 ноября 2024 года. Екатеринбург : Уральский государственный лесотехнический университет, 2024. 323 с. С. 89–96. ISBN: 978-5-94984-931-6.
6. Кулакова 2022 — *Кулакова Л. И.* Инновационные риски применения искусственного интеллекта в предпринимательских структурах. EDN: PGMWXQ // Вестник академии знаний. 2022. 50 (3): 180-185. ISSN: 2304-6139; eISSN: 2687-0983.
7. Харитонова 2025 — *Харитонова Ю. С.* Правовые подходы к регулированию искусственного интеллекта: уроки китайского права. DOI: 10.17803/1729-5920.2025.222.5.130-138. EDN: BKCABK // Lex Russica. 2025; 78(5):130-138. ISSN: 1729-5920; eISSN: 2686-7869.
8. Шейкин 2024 — *Шейкин А. Г.* Принципы законодательного регулирования искусственного интеллекта в США и их влияние на развитие технологического сектора. DOI: 10.21639/2313-6715.2024.2.3. EDN: LZPRWH // Пролог: журнал о праве = Prologue: Law Journal. 2024; 2:28–38. eISSN: 2313-6715.
9. Jobin 2019 — *Jobin A., Ienca M., Vayena E.* The global landscape of AI ethics guidelines. DOI:10.1038/s42256-019-0088-2 // Nature Machine Intelligence. 2019; 1(2):389–399.
10. Krause 2025 — *Krause D., Krause E.* Auditing GenAI Systems: Ensuring Responsible Deployment // The Capso Institute Journal of Financial Transformation. Technology. 60th ed. 2025. Pp. 20-27. Текст : электронный. URL: https://www.researchgate.net/publication/388879215_Auditing_GenAI_Systems_Ensuring_Responsible_Deployment (дата обращения: 01.05.2025)
11. Lin 2025 — *Lin T.* Enterprise AI Governance Frameworks: A Product Management Approach to Balancing Innovation and Risk. DOI: 10.56726/IRJMETS67008. 2025. Текст : электронный. URL: https://www.researchgate.net/publication/390145149_ENTERPRISE_AI_GOVERNANCE_FRAMEWORKS_A_PRODUCT_MANAGEMENT_APPROACH_TO_BALANCING_INNOVATION_AND_RISK (дата обращения: 01.05.2025).
12. Mökander 2023 — *Mökander J., Schuett J., Kirk H.R., & Floridi L.,* Auditing large language models: a three-layered approach. DOI:10.1007/s43681-023-00289-2. AI Ethics. 202). 4(4):1085-1115.

References

1. Averkin A. N. Ob»yasnitel'nyy iskusstvennyy intellekt v bol'shikh rechevykh modelyakh [Explanatory Artificial Intelligence in Large Speech Models]. EDN: ROFNZG. *Rechevyye tekhnologii*. 2024; 1: 3–13. ISSN: 2305-8129; eISSN: 2305-8137 (in Russ.).
2. Bylevsky P. G. Kul'turologicheskaya dekonstruktsiya sotsial'nokul'turnykh ugroz ChatGPT informatsionnoy bezopasnosti rossiyskikh grazhdan [Culturological Deconstruction of Socio-Cultural Threats ChatGPT to the Information Security of Russian Citizens]. DOI: 10.7256/2454-0757.2023.8.43909. EDN: UZDRFW. *Philosophy and Culture.* 2023; 8: 46–56. ISSN: 1999-2793; eISSN: 2454-0757 (in Russ.).
3. Volobuev A. V. Etika iskusstvennogo intellekta, diskriminatsiya i neravenstvo [Ethics of Artificial Intelligence, Discrimination and Inequality]. DOI: 10.30884/vglob/2023.03.04. EDN: DNUTN. *Vek globalizatsii*. 2023; 3: 48-62. ISSN: 1994-9065 (in Russ.).
4. Golovina T. A. Upravleniye riskami organizatsiy v usloviyakh tsifrovoy ekonomiki Risk management of organizations in the digital economy. By T. A. Golovina, I. L. Avdeeva, D. A. Sukhanov. DOI: 10.24412/2304-6139-2022-48-1-55-61. EDN: CXBFBR. *Vestnik akademii znaniy*. 2022; 48(1): 55–61. ISSN: 2304-6139; eISSN: 2687-0983 (in Russ.).
5. Kiseleva L. A. Pravovoye regulirovaniye iskusstvennogo intellekta: ot «chernogo yashchika» k prozrachnosti algoritmov [Legal regulation of artificial intelligence: from the "black box" to the transparency of algorithms]. By L. A. Kiseleva, A. S. Shaidurov. EDN: LONHBK. *Tsivilizatsionnyye peremeny v Rossii* : materials of the 14th All-Russian scientific and practical conference, Yekaterinburg, November 29 2024. Yekaterinburg : Ural State Forest Engineering University Publ., 2024. 323 c. Pp. 89–96. ISBN: 978-5-94984-931-6 (in Russ.).
6. Kulakova L. I. Innovatsionnyye riski primeneniya iskusstvennogo intellekta v predprinimatel'skikh strukturakh [Innovative risks of using artificial intelligence in business structures]. EDN: PGMWXQ. *Vestnik akademii znaniy*. 2022. 50 (3): 180–185. ISSN: 2304-6139; eISSN: 2687-0983 (in Russ.).
7. Kharitonova Yu. S. Pravovyye podkhody k regulirovaniyu iskusstvennogo intellekta: uroki kitayskogo prava [Legal approaches to regulating artificial intelligence: lessons

from Chinese law][1]. DOI: 10.17803/1729-5920.2025.222.5.130-138. EDN: BKCABK. *Lex Russica.* 2025; 78(5):130–138. ISSN: 1729-5920; eISSN: 2686-7869 (in Russ.).

8. Sheikin A. G. Printsipy zakonodatel'nogo regulirovaniya iskusstvennogo intellekta v SSHA i ikh vliyaniye na razvitiye tekhnologicheskogo sektora [Principles of legislative regulation of artificial intelligence in the USA and their impact on the development of the technology sector]. DOI: 10.21639/2313-6715.2024.2.3. EDN: LZPRWH. *Prologue: Law Journal.* 2024; 2:28–38. eISSN: 2313-6715 (in Russ.).
9. Jobin A., Ienca M., Vayena E. The global landscape of AI ethics guidelines. DOI:10.1038/s42256-019-0088-2. *Nature Machine Intelligence*. 2019; 1(2):389–399..
10. Krause D., Krause E. Auditing GenAI Systems: Ensuring Responsible Deployment. *The Capso Institute Journal of Financial Transformation. Technology.* 60th ed. 2025. Pp. 20-27. Text : electronic. URL: https://www.researchgate.net/publication/388879215_Auditing_GenAI_Systems_Ensuring_Responsible_Deployment (accessed: 05/05/2025).
11. Lin T. Enterprise AI Governance Frameworks: A Product Management Approach to Balancing Innovation and Risk. DOI: 10.56726/IRJMETS67008. 2025. Text : electronic. URL: https://www.researchgate.net/publication/390145149_ENTERPRISE_AI_GOVERNANCE_FRAMEWORKS_A_PRODUCT_MANAGEMENT_APPROACH_TO_BALANCING_INNOVATION_AND_RISK (accessed: 05/01/2025).
12. Mökander J., Schuett J., Kirk H. R., & Floridi L. Auditing large language models: a three-layered approach. DOI:10.1007/s43681-023-00289-2. *AI Ethics*. 202). 4(4):1085-1115.

Информация об авторах:

Островская Анна Александровна — кандидат экономических наук, Высшая школа управления РУДН, SPIN-код: 4287-0228, AuthorID: 307787; **Денисова София Денисовна, Машинистов Виктор Евгеньевич, Оленев Александр Юрьевич**— магистранты.

Место работы авторов: федеральное государственное автономное образовательное учреждение высшего образования «Российский университет дружбы народов имени Патриса Лумумбы», ВШУ, ул. Миклухо-Маклая, 6, Москва, 117198, Россия.

Information about the authors:

Ostrovskaya Anna A. — Candidate of Economic Sciences, Higher School of Management, RUDN University; **Denisova Sofia D., Mashinistov Victor E., Olenev Alexander Yu.** — Master's students.
Authors' place of work: Peoples' Friendship University of Russia named after Patrice Lumumba», Higher School of Management, 6 Miklukho-Maklaya St., Moscow, 117198, Russia.

Статья поступила в редакцию 05.05.2025; одобрена после рецензирования 23.05.2025; принята к публикации 27.06.2025.
The article was submitted 05/05/2025; approved after reviewing 05/23/2025; accepted for publication 06/27/2025.

1