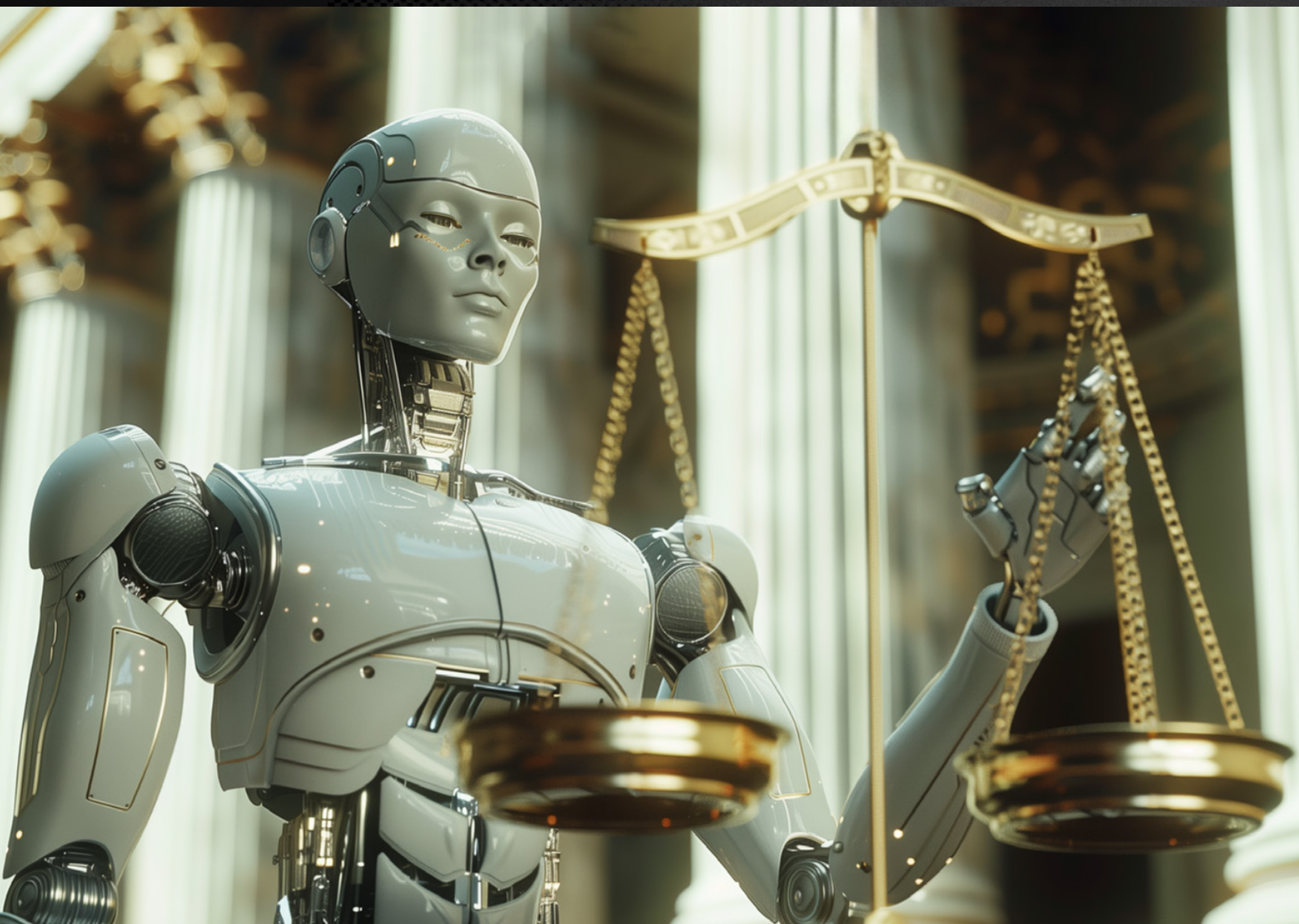
Яков и Партнёры × АЛЬЯНС В СФЕРЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Регулирование генеративного ИИ: правовой анализ и риски для РФ

Максим Болотских, Никита Абанитов, Никита Власов, Алина Гилязова

Москва, 2024



Содержание

Источник: открытые источники,
анализ «Яков и Партнёры»

5 наиболее актуальных для России рисков

генеративного ИИ выявлены экспертами «Яков и Партнёры»

Введение. Поиск баланса между регулированием генеративного искусственного интеллекта и свободой для развития этой технологии

Генеративный искусственный интеллект (ГИИ) – это разновидность ИИ на базе больших языковых моделей, позволяющих создавать новые данные различных модальностей на основе информации, которая использовалась для их обучения.

К таким моделям относятся, в частности, GPT-4 (OpenAI), Llama 3 (Meta*), YandexGPT («Яндекс»), GigaChat (Сбер). Генеративный ИИ может создавать текст, программный код, высококачественные изображения, а также короткие видео.

Как сам ГИИ, так и сервисы на его основе развиваются беспрецедентно высокими темпами. Исследование «Искусственный интеллект в России – 2023: тренды и перспективы», которое провела компания «Яков и Партнёры», показало: к 2028 г. полный экономический потенциал генеративного ИИ в России может достичь 0,8–1,3 трлн руб.

Крупнейшие компании уже внедряют ГИИ в собственные процессы и продукты. У этих моделей широкий спектр областей применения – от поддержки клиентов и обеспечения работы бэк-офиса до моделирования молекулярных взаимодействий и прогнозирования решений о кандидатах на получение статуса лекарственного препарата (MIT-IBM Watson AI Lab, 2023)[1].

Быстрое развитие технологии ГИИ стало возможно благодаря отсутствию жестких ограничений. Вместе с тем новые технологии формируют новые риски в области общественной безопасности, такие как предоставление ложных и неэтичных ответов, мошенничество и т. д. Ответы на эти новые вызовы отрасль должна искать совместно с обществом и государством.

* Организация, деятельность которой запрещена на территории Российской Федерации.

В отсутствие общепризнанных подходов к нивелированию рисков каждая страна ищет баланс между регулированием и свободой самостоятельно

Возможность реализации рисков подтверждают первые широко известные инциденты, например случай, произошедший в июне 2023 г. в США: адвокат использовал ГИИ для сбора сведений о прецедентах, доказывающих правоту его клиента, но не знал, что сервис может сгенерировать ложную информацию. В результате он предъявил суду несуществующие дела[2]. Еще один случай произошел в феврале 2024 г. в Китае: сотрудника финансовой службы транснациональной компании обманом заставили выплатить 25 млн долл. США мошенникам, которые выдавали себя за финансового директора этой организации, используя технологию дипфейк (создание поддельных изображений, видеороликов и других материалов с помощью генеративного ИИ)[3].

В отсутствие общепризнанных подходов к нивелированию рисков каждая страна ищет баланс между регулированием и свободой самостоятельно. Именно поэтому возникает все больше дискуссий о таких правилах работы ГИИ, которые помогут не только сохранять благоприятные условия для развития этой технологии и привлечения инвестиций, но и обеспечивать общественную безопасность.

По этим причинам при подготовке настоящего исследования мы ставили две главные цели: во-первых, проанализировать нормативно-правовую базу разных стран, связанную с ГИИ, а во-вторых, выявить риски, которые наиболее значимы для российского рынка, и предложить способы снижения их влияния.

Результаты этой работы позволят комплексно проанализировать существующие подходы к регулированию, а также критически и всесторонне рассмотреть ключевые риски, чтобы в дальнейшем разработать методы снижения их влияния в России.

Правовой ландшафт в области генеративного ИИ

Основные нормативно-правовые акты и динамика принятия законопроектов

На данный момент существуют три подхода к формированию нормативно-правовой базы для регулирования деятельности, связанной с ГИИ:

1 **Принятие общих норм регулирования сферы информационных технологий (ИТ),** например законов о персональных данных, которые охватывают среди прочего и технологию ИИ (некоторые страны, в частности США и Великобритания, включили в такие акты отдельные положения, касающиеся ИИ).

---

2 **Принятие комплексных нормативных документов для регулирования ИИ в целом,** включающих среди прочего отдельные положения о ГИИ. Например, это описанные далее закон ЕС и законопроект, который рассматривается в Бразилии.

---

3 **Принятие отдельных нормативно-правовых актов для ГИИ,** таких как временные меры Китая или соглашение между компаниями в США.

Несмотря на то что законы, регулирующие генеративный ИИ, принимаются все активнее, международная правовая база в этой сфере пока лишь формируется, а страны тестируют принятые меры и следят за тем, насколько эффективно те помогают предотвращать инциденты.

Евросоюз

В Евросоюзе принят первый всеобъемлющий нормативно-правовой акт о регулировании ИИ. 21 мая 2024 г. Совет Европейского союза одобрил Закон ЕС об искусственном интеллекте[4] – комплексный документ, определяющий конкретные меры регулирования в зависимости от уровня риска, связанного с использованием ИИ, в частности генеративного ИИ.

Бразилия

Законопроект № 2338/2023, который с 2023 г. рассматривается в Бразилии, похож на закон ЕС. В основе мер регулирования лежит рискориентированный подход. Перечень областей регулирования адаптирован к особенностям местного рынка, а степень детализации требований существенно ниже, чем в законе ЕС.

Китай

Китай в 2023 г. принял наиболее детальные временные меры по регулированию ГИИ – Требования к безопасности сервисов генеративного ИИ. Хотя это временный документ и формулировки еще могут быть уточнены, изложенные в нем требования обязательны для исполнения всеми разработчиками данной технологии.

США и Канада

В США и Канаде применяется механизм саморегулирования, который заключается в том, что крупнейшие компании – разработчики инструментов ИИ дают гарантии развивать эту технологию и обязуются управлять соответствующими рисками, а также способствовать исследованиям в целях поиска безопасных решений в сфере ИИ и обеспечивать плодотворное сотрудничество между представителями отрасли и законодательной властью.

В июле 2023 г. семь компаний, ведущих разработки в сфере ИИ (Amazon, Anthropic, Google, Inflection, Meta*, Microsoft и OpenAI), подписали документ «Развитие безопасного, защищенного и заслуживающего доверия искусственного интеллекта» и взяли на себя обязательство самостоятельно регулировать разработки в области ИИ.

* Организация, деятельность которой запрещена на территории Российской Федерации.

Россия

В России нормативная база для регулирования ИТ-отрасли содержит документы, касающиеся функционирования ИИ. В частности, это общие нормативно-правовые акты в отношении ИИ, охватывающие и генеративный ИИ. В их числе Национальная стратегия развития искусственного интеллекта на период до 2030 г. и Концепция развития регулирования отношений в сфере технологий ИИ и робототехники до 2024 г.

Для ГИИ также применяется отдельный механизм саморегулирования. Это общая тенденция в сфере законодательства, позволяющая сохранять возможности для развития данной технологии. 13 марта 2024 г. участники Альянса в сфере искусственного интеллекта (включая Сбер, «Яндекс», MTS AI) подписали Декларацию об ответственной разработке и использовании сервисов на основе генеративного ИИ. Документ определяет этические принципы и рекомендации, способствующие ответственному отношению к ИИ. Декларация развивает и детализирует опубликованный в 2021 г. Кодекс этики в сфере искусственного интеллекта, касающийся ГИИ[5].

Великобритания, Сингапур, Южная Корея, Япония, Израиль и Казахстан

В Великобритании, Сингапуре, Южной Корее, Японии, Израиле и Казахстане нормативно-правовая база представлена законопроектами, которые определяют общие требования, включая этические нормы, а также требования создать среду для развития ИИ, обеспечивать его безопасное использование и т. д. В законодательстве Японии отдельно указано, что для обучения больших языковых моделей генеративного ИИ можно использовать любой контент[6].

Объединенные Арабские Эмираты

Объединенные Арабские Эмираты не принимали мер в области регулирования с целью обеспечить максимально свободное развитие этой технологии.

Во всех странах интенсивно развивается нормативно-правовая база в области генеративного ИИ (1/3)

Примеры ключевых документов (неисчерпывающий список)

● Законы, временные меры, президентские указы – юридически обязательные для исполнения нормы

● Кодексы или декларации – добровольные соглашения нескольких компаний (самостоятельное регулирование)

○ Законопроекты и прочее – документы, находящиеся на этапе обсуждения

Страна	2019	2020	2021	2022	2023	2024
	Национальная стратегия развития искусственного интеллекта на период до 2030 г. (утверждена Указом Президента Российской Федерации № 490) ●	Перспективная программа стандартизации по приоритетному направлению «Искусственный интеллект» на 2021–2024 гг. ● Концепция развития регулирования отношений в сфере технологий ИИ и робототехники до 2024 г. ○ Перечень поручений Президента РФ по итогам конференции «Путешествие в мир искусственного интеллекта» ● Федеральный закон № 258-ФЗ «Об экспериментальных правовых режимах в сфере цифровых инноваций в Российской Федерации» ● Федеральный закон № 123-ФЗ «О проведении эксперимента по установлению специального регулирования в целях создания необходимых условий для разработки и внедрения технологий искусственного интеллекта в... городе Москве» ●	Кодекс этики в сфере ИИ ● Перечень поручений Президента РФ по итогам конференции «Путешествие в мир искусственного интеллекта» ●	Перечень поручений Президента РФ по итогам конференции «Путешествие в мир искусственного интеллекта» ●	Перечень поручений Президента РФ по итогам конференции «Путешествие в мир искусственного интеллекта» ●	Декларация об ответственной разработке и использовании сервисов на основе генеративного ИИ (подписана, в частности, Сбером, «Яндексом» и MTS AI в 2024 г.) ● Поправки к Национальной стратегии развития искусственного интеллекта на период до 2030 г. ●
				Положение об администрировании информационных услуг глубокого синтеза ●	Требования к безопасности сервисов генеративного ИИ (временные меры) ● Проект мер по этической экспертизе (в сфере НИОКР) ●	
	Отсутствует законодательная база					

Во всех странах интенсивно развивается нормативно-правовая база в области генеративного ИИ (2/3)

Примеры ключевых документов (неисчерпывающий список)

● Законы, временные меры, президентские указы – юридически обязательные для исполнения нормы

● Кодексы или декларации – добровольные соглашения нескольких компаний (самостоятельное регулирование)

○ Законопроекты и прочее – документы, находящиеся на этапе обсуждения

Страна	2019	2020	2021	2022	2023	2024
США			● Закон о найме, направленный против предвзятости, связанной с использованием ИИ (LL 144)	● Закон о подотчетности алгоритмов	● Указ Байдена о безопасном, защищенном и заслуживающем доверия ИИ ● Документ «Развитие безопасного, защищенного и заслуживающего доверия ИИ» (обязательства семи компаний по самостоятельному регулированию ИИ) ○ Руководство по регистрации авторских прав: произведения, содержащие материалы, созданные ИИ ○ Концепция NIST по управлению рисками, связанными с ИИ ○ Закон об ответственности за дипфейки	
ЕС	○ Руководящие этические принципы для заслуживающего доверия ИИ		○ Предложен проект Закона ЕС об искусственном интеллекте		● Директива ЕС об ответственности за ИИ (режим урегулирования ущерба, причиненного ИИ)	● Закон ЕС об искусственном интеллекте
Бразилия	○ Законопроект, устанавливающий принципы использования ИИ	○ Законопроект, содержащий основополагающие и руководящие указания по разработке и применению ИИ	○ Законопроект, предусматривающий этические рамки и руководящие принципы, регламентирующие разработку и использование ИИ		○ Законопроект № 2338/2023, объединяющий более ранние предложения за 2019–2021 гг. (на рассмотрении с 2023 г.)	
Великобритания				○ Документ о политике регулирования ИИ ● Закон об авторском праве, дизайне и патентах от 1988 г. (последняя поправка внесена в 2022 г.)	○ Инновационный подход к регулированию ИИ ● Закон о безопасности в интернете	

Во всех странах интенсивно развивается нормативно-правовая база в области генеративного ИИ (3/3)

Примеры ключевых документов (неисчерпывающий список)

● Законы, временные меры, президентские указы – юридически обязательные для исполнения нормы

● Кодексы или декларации – добровольные соглашения нескольких компаний (самостоятельное регулирование)

○ Законопроекты и прочее – документы, находящиеся на этапе обсуждения

Страна	2019	2020	2021	2022	2023	2024
	○ Типовая концепция управления искусственным интеллектом, разработанная PDPC, первая версия	○ Типовая концепция управления искусственным интеллектом, разработанная PDPC, вторая версия				
			○ Стратегия реализации заслуживающего доверия ИИ ● Указ о введении в действие основополагающего закона об интеллектуальной информатизации		○ Законопроект о развитии индустрии ИИ и создании фонда для развития заслуживающего доверия ИИ ○ Законопроект об ответственности за искусственный интеллект	
	○ Социальные принципы человекоориентированного ИИ		● Закон об авторском праве (принят в 1899 г., последняя поправка внесена в 2021 г.) ○ Отчет «Управление искусственным интеллектом в Японии, версия 1.1»		○ Книга «Подход Японии к регулированию ИИ и его влияние на председательство в G7 в 2023 г.»	
						○ Концепция развития искусственного интеллекта на 2024–2029 гг. (на рассмотрении)
		● Совместное заявление членов-основателей глобального партнерства по ИИ		○ Законопроект об искусственном интеллекте и данных ○ Комплекс законопроектов, призванных укрепить защиту частной жизни канадцев и доверие к цифровой экономике	● Добровольный кодекс поведения по ответственной разработке передовых систем генеративного ИИ и управлению ими	
					○ Политика Израиля в области регулирования и этики искусственного интеллекта	

Подход к анализу на основе жизненного цикла модели

Требования и рекомендации для различных заинтересованных лиц, изложенные в нормативных актах, зачастую смешиваются и таким образом создают путаницу, а многие документы, касающиеся генеративного ИИ, структурированы непоследовательно. Поэтому для комплексного анализа мер в данной сфере требовался подход, позволяющий сформировать гармонизированное представление о мировой практике.

В связи с этим был выбран подход, в рамках которого учитывается жизненный цикл сервиса на основе ГИИ. Проводить анализ при помощи этой методики целесообразно по двум причинам. Во-первых, требования, описанные в законах, применяются для снижения рисков. В то же время риски могут возникнуть на каждом из этапов жизненного цикла – от проектирования продукта до использования ответа сервиса пользователем. Во-вторых, большинство мер по смягчению последствий использования ГИИ носят технический характер и могут быть приняты на протяжении всего жизненного цикла сервиса.

Жизненный цикл сервиса на основе генеративного ИИ

Анализ правовой базы с учетом жизненного цикла

Нормативно-правовая база не охватывает весь жизненный цикл сервиса на основе ГИИ ни в одной из стран

На текущий момент нормативно-правовая база не охватывает весь жизненный цикл сервиса на основе ГИИ ни в одной из стран. Наиболее широкий охват имеют меры, принятые в Китае. Например, для тестирования модели необходимо не менее 4 тыс. вопросов для оценки безопасности и качества и не менее 100 вопросов для проверки на «отказ отвечать» с 95-процентным порогом качества. Также необходимо обеспечить разнообразие данных и иметь более одного источника данных для каждого языка (китайского, английского и т. д.) и для каждого типа данных, таких как текст, изображения, видео и т. д., при этом соотношение между отечественными и зарубежными источниками должно быть разумным. Что касается маркировки данных, в документах не только описано, как нужно ее осуществлять, но и указано, что для этого необходимо выделить персонал, провести его оценку и присвоить соответствующую квалификацию, разделить функции маркировщика на маркировку и рецензирование данных, а также позаботиться о том, чтобы в рамках одной задачи по маркировке один и тот же маркировщик не выполнял более одной функции.

В остальных странах меры, связанные с ИИ, в основном описаны в качестве общих требований. В ЕС и в Бразилии используется рискориентированный подход и детально описаны потенциальные области применения с учетом величины риска: неприемлемый, высокий и т. д. Неприемлемый риск определяется для тех областей, в которых применять системы ИИ запрещено. Это сферы, где существуют следующие риски: побуждение к опасному для здоровья поведению, нелегитимная оценка, социальная классификация. Высокий риск определяется для тех областей, в которых предусмотрены дополнительные меры регулирования, такие как требование уведомлять компетентный орган о запуске сервиса, создающего высокий риск, и регистрировать такой сервис в государственной базе. К подобным областям относятся образование, управление персоналом, государственные услуги и т. д.

Для систем, создающих высокий риск, на этапе тестирования описаны правила ведения журналов. В частности, в журнале указывают дату и время начала и окончания каждого сеанса использования системы. Помимо этого, ведется справочная база данных, по которой система проверяет входные данные.

Существующие законы не охватывают всего жизненного цикла моделей – они в основном определяют общие требования

○ Отсутствуют требования
⊕ Есть детальные требования
⊕ Есть общие требования

Этапы жизненного цикла	Шаги	🇷🇺	🇨🇳	🇺🇸	🇪🇺	🇧🇷	🇬🇧	🇸🇬	🇰🇷	🇯🇵	🇰🇿	🇮🇱	🇨🇦
1 Разработка технологии и проектирование продуктов	Разработка технологии, внедрение инноваций и проектирование продуктов	○	⊕	⊕	⊕	⊕	⊕	○	⊕	⊕	⊕	○	⊕
2 Обучение и постоянное дообучение моделей	Сбор и подготовка данных для обучения/ дообучения моделей Наиболее раскрытая область	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	○	⊕
	Обучение/ дообучение и оценка качества моделей	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
3 Промышленное развертывание сервиса	Развертывание сервиса и мониторинг стабильности его работы	○	○	⊕	⊕	○	○	⊕	○	○	○	○	⊕
4 Работа сервиса	Запрос к сервису Наименее раскрытая область	○	⊕	○	○	○	○	○	○	○	○	○	○
	Формирование и выдача ответа сервисом	⊕	⊕	⊕	⊕	○	⊕	○	⊕	⊕	○	○	⊕
	Использование ответа сервиса Наиболее раскрытая область	⊕	⊕	⊕[1]	⊕	⊕	⊕	⊕	⊕	⊕	○	○	⊕

1. Активно обсуждаются правила установления авторских прав, но конкретных законов пока не принято. Например, если пользователь дал генеративному ИИ подсказку, то право собственности принадлежит этому человеку (при условии что он сделал нечто большее, чем просто направил подсказку)

Источник: анализ нормативно-правовой базы различных стран, анализ «Яков и Партнёры»

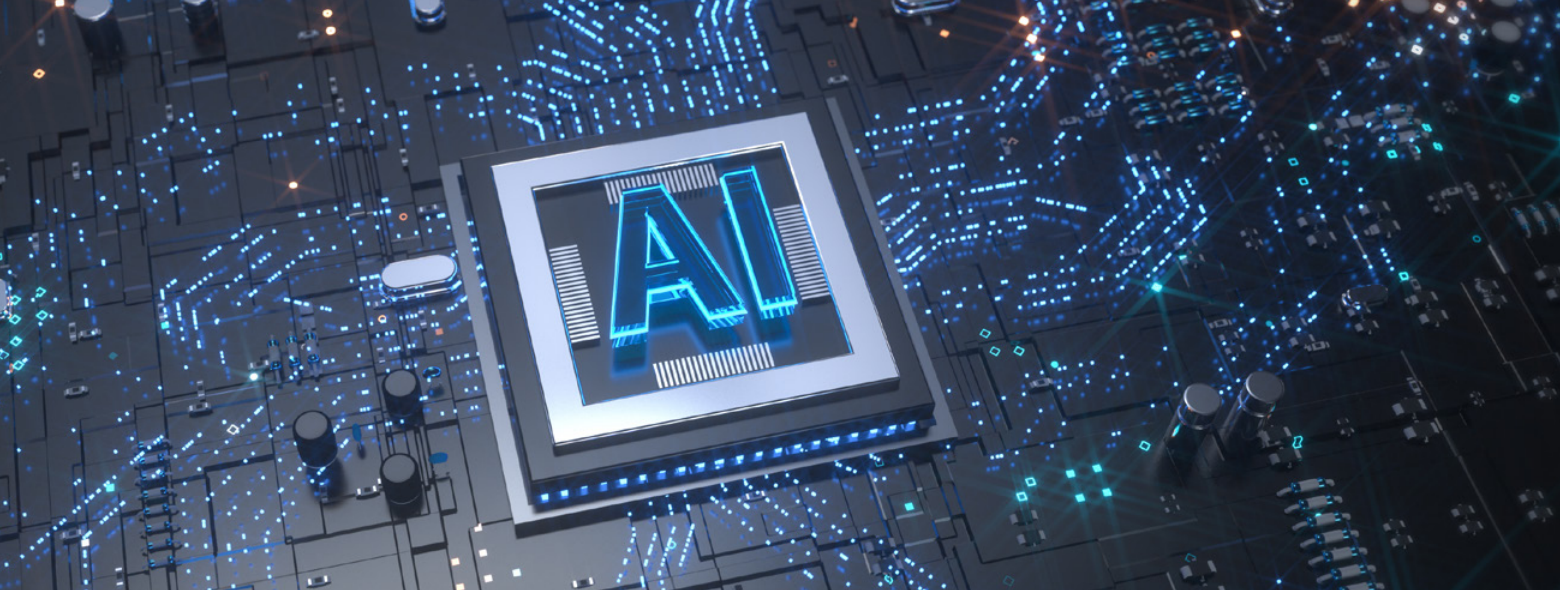
С одной стороны, акцент на общие требования создает сложности для разработчиков технологии и сервисов, поскольку не всегда ясно, как именно выполнять введенные нормы. С другой стороны, такой подход создает возможность для «мягкого» регулирования, так как позволяет разработчикам самостоятельно адаптировать требования.

Например, существует следующее требование:

> «Осуществляйте маркировку (видимую или скрытую, в зависимости от целесообразности и технической возможности) информации, создаваемой с помощью сервисов на основе генеративного искусственного интеллекта, там, где это разумно применимо и оправданно с учетом специфики сервиса[7]»

Данная формулировка не определяет, как именно осуществлять маркировку, но это позволяет разработчикам самим выбирать способы маркирования, его формат и сервисы, в которых оно необходимо, исходя из потенциальных рисков и произошедших инцидентов.

В дальнейшем, по мере развития технологии, технологические компании протестируют различные подходы и определят наиболее эффективные из них. Тогда требования могут быть уточнены, детализированы и включены в нормативные документы.



> Необходима совместная работа регулирующих органов и технологических компаний для выработки мер по регулированию генеративного ИИ. Временные меры в Китае были разработаны именно так: в их подготовке участвовали мы и другие технологические фирмы

Заместитель директора Huawei по взаимодействию с государственными органами

«Жесткий» режим регулирования применяют в тех случаях, когда отдельный этап жизненного цикла сервиса создает значительный риск. Тогда власти страны вводят детальные, трудновыполнимые требования, такие как исключение определенных источников, лицензирование и т. д.

> Европейские требования самые сложные для соблюдения поставщиками. Следом идут требования Китая, введенные для внешних систем, и уже потом – США

Член комитета по регулированию ИИ в Европе

Национальные меры регулирования можно представить с помощью градиента от «мягких» к «жестким» (1/4)

Пример требований по этапам жизненного цикла (неисчерпывающий список)

«Мягкое» регулирование — «Жесткое» регулирование

Страна, применяющая подход

Жизненный цикл сервиса	Способ регулирования / подход 01	02	03
Этап 1. Разработка технологии, внедрение инноваций и проектирование сервисов			
Внедрение инноваций	**Общие требования с детализацией, например:** — Создать новые рабочие места — Стимулировать частные инвестиции в ИИ — Поддерживать справедливую конкурентную среду		
Наличие классификации сервисов на основе профиля рисков	Прочие страны **Отсутствие классификации по сферам/областям** Требования, изложенные в законе, применимы ко всем областям	**Выделены две группы, исходя из профиля рисков** 1. Высокий риск: могут оказывать значительное влияние на жизнь, физическую безопасность и защиту основных прав людей Мера: могут быть запрещены, если государственный орган выявит риски 2. Прочие области	**Классификация с перечнем из 3–4 областей** 1. Неприемлемый риск: области, где есть риски социальной классификации, вреда для здоровья, нелегитимной оценки Мера: запрещены к использованию 2. Высокий риск: образование, здравоохранение, кредитный скоринг Мера: дополнительные меры регулирования (регистрация систем и т. д.) 3. Ограниченный риск: сервисы на основе генеративного ИИ 4. Минимальный риск: прочие системы

Национальные меры регулирования можно представить с помощью градиента от «мягких» к «жестким» (2/4)

Пример требований по этапам жизненного цикла (неисчерпывающий список)

«Мягкое» регулирование — «Жесткое» регулирование

Страна, применяющая подход

Жизненный цикл сервиса	Способ регулирования / подход 01	02	03
Этап 2. Обучение и постоянное дообучение моделей			
Сбор и подготовка данных для обучения и постоянного дообучения моделей	Прочие страны **Общие требования** Данные должны быть качественными и актуальными, не содержать ошибок и т. д.	**Необходимость собирать историю данных** Данные должны быть качественными. Необходимо знать об их происхождении: откуда они взяты, как собирались, как передавались и т. д.	**Исключение отдельных источников** Не использовать данные из источников, входящих в черный список Обеспечить разумное соотношение между **отечественными и зарубежными источниками**
Обучение и постоянное дообучение моделей, оценка качества обучения	Прочие страны **Общие требования к тестированию качества обучения моделей** **Дополнительно:** — Рекомендуется проводить аудит систем силами сторонних экспертов — Национальный институт стандартов и технологий устанавливает добровольные стандарты для всестороннего тестирования моделей ИИ		**Детальные требования к качеству и оценке** — Для любой модели необходима оценка по 31 риску — Перечень рисков указан в приложении к закону — Детальные требования к тестированию моделей и качеству с указанием необходимого показателя качества и количества тестовых вопросов Например: не менее 4 тыс. тестовых вопросов с показателем качества на уровне 95%

Национальные меры регулирования можно представить с помощью градиента от «мягких» к «жестким» (3/4)

Пример требований по этапам жизненного цикла (неисчерпывающий список)

«Мягкое» регулирование — «Жесткое» регулирование

Страна, применяющая подход

Жизненный цикл сервиса	Способ регулирования / подход 01	02	03
Этап 3. Развертывание сервисов и мониторинг стабильности их работы	Прочие страны		
	Общие требования к оценке стабильности работы сервисов и внедрению процедур системного мониторинга **Отдельные требования к госорганам по созданию кибербезопасности** Создание программы по обеспечению кибербезопасности при разработке инструментов ИИ. Задача такой программы – искать и устранять уязвимости в критически важном программном обеспечении	**Создание механизмов человеческого надзора и мониторинга для систем, создающих высокий риск, ведение соответствующей документации**	**Для использования модели генеративного ИИ в сервисе необходимо получить лицензию:** — В Китае – для всех сервисов с генеративным ИИ — В ЕС – только для сервисов, применяемых в областях с высоким риском **Ведение журналов для систем, создающих высокий риск, в ЕС** — Запись каждого сеанса использования системы (точное время начала и окончания каждого сеанса) — Ведение справочной базы данных, по которой система проверяет входящие данные — Идентификация физических лиц, участвующих в проверке результатов Автоматически сгенерированные журналы, содержащие информацию о работе системы, должны храниться у поставщика не менее 6 месяцев

Национальные меры регулирования можно представить с помощью градиента от «мягких» к «жестким» (4/4)

Пример требований по этапам жизненного цикла (неисчерпывающий список)

«Мягкое» регулирование — «Жесткое» регулирование

Страна, применяющая подход

Жизненный цикл сервиса	Способ регулирования / подход 01	02	03
Этап 4. Работа сервиса			
Запрос к сервису			**Проверка данных/ промпта на безопасность** В каждом диалоге вводимые пользователем данные должны проверяться на безопасность, а сервис должен быть ориентирован на создание положительного контента
Формирование и выдача ответа сервисом	Прочие страны **Общие требования** Сервисы на базе ГИИ: — Не должны давать дискриминирующих результатов — Должны обеспечивать соблюдение законов и морально-этических норм	**Общие требования и требования к маркировке** Необходимо маркировать контент (видео, изображения и т. д.) с помощью водяных знаков Необходимо информировать пользователей о взаимодействии с контентом, созданным ИИ (аудио)	**Общие требования и детальные требования к маркировке для каждого типа контента, включая аудио, видео и изображения** Например, для аудиоконтента: не менее одного устного и письменного заявления о том, что контент сгенерирован ИИ
Использование ответа сервиса	**Информирование пользователя об ответственности и рисках** Требования к информированию пользователей: — В пользовательском соглашении сервиса — В интерфейсе сервиса, например в виде дисклеймеров, уведомлений, примечаний Также пользователи должны быть проинформированы о рисках, связанных с генеративным ИИ		

Матрица правового ландшафта

Чтобы сформировать гармоничное представление о правовом ландшафте в области ГИИ, а также дать ему комплексную оценку, компания «Яков и Партнёры» разработала матрицу, которая отражает зависимость между степенью покрытия рисков, связанных с генеративным ИИ, и степенью свободы, предоставляемой технологическим компаниям.

Матрица включает в себя две оси, которые описаны далее.

1

Ось OX: степень свободы для компаний-разработчиков

Этот показатель характеризует то, насколько легко технологическим компаниям выполнять требования регулирующего органа.

— **Высокая степень свободы.**
Поставщикам легко выполнить требования (например, общие требования к сервисам: этические требования, необходимость реагировать на инциденты и т. д.).

— **Средняя степень свободы.**
Регулирование создает значимые ограничения для разработки технологии (например, необходимо проводить тестирование по определенным критериям и предотвращать использование данных из источников, внесенных в черный список).

— **Низкая степень свободы.**
Регулирование создает существенные препятствия для развития технологии (например, требование вести журнал, в который автоматически записываются любые действия системы [ЕС]; требование проходить аудит и получать лицензию на запуск системы).

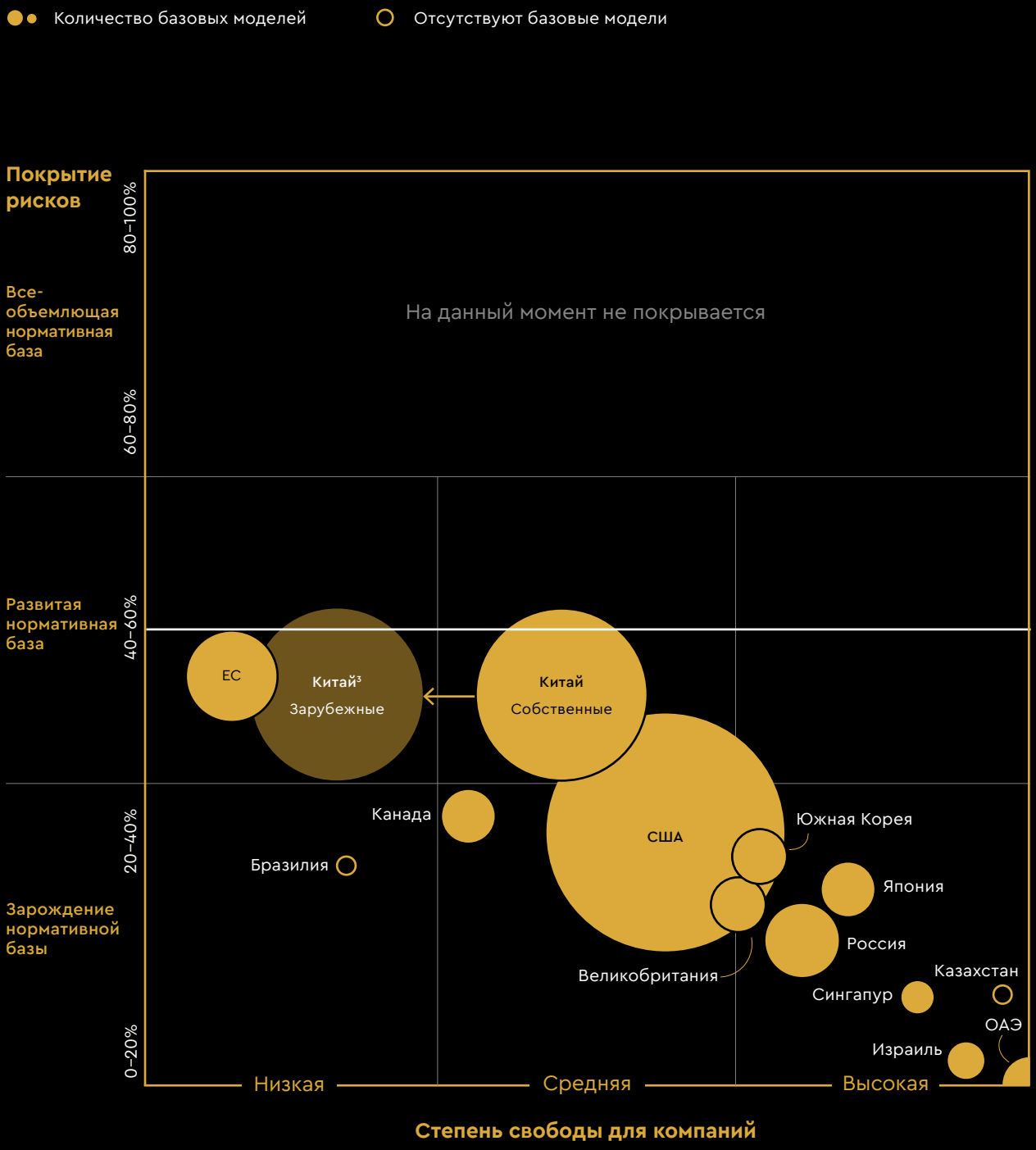
2

Ось OY: степень покрытия рисков

Этот показатель отражает процентную долю рисков, охваченных требованиями к генеративному ИИ в правовой базе страны, от общего количества рисков, выявленных компанией «Яков и Партнёры» на всех этапах жизненного цикла сервиса.
В зависимости от значения данного показателя выделено три категории: зарождение нормативной базы, развитая нормативная база, всеобъемлющая нормативная база.

Правовой ландшафт в области генеративного ИИ

Размер круга соответствует количеству базовых моделей ГИИ.
Иллюстративно



1. Закон ЕС об искусственном интеллекте в основном затрагивает регулирование систем ИИ, создающих высокий риск
2. Доля китайских разработок на рынке моделей LLM составляет 40%, доля моделей из США – 50% (2022 г.). Источник: Reuters, Bloomberg
3. Внешним поставщикам LLM трудно выполнять такие требования, как недопущение использования сведений из черного списка при формировании массива данных, выполнение предписаний цензуры и т. д.

США, Израиль, Сингапур и ОАЭ

В США, Израиле, Сингапуре, ОАЭ созданы благоприятные условия: предоставлена высокая степень свободы, помогающая развивать искусственный интеллект и привлекать инвестиции.

Россия

В России соблюдается баланс между жесткостью правил и свободой для развития технологий.

Китай

В Китае используется две модели:

1. **Для разработчиков внутри страны.**
   Создаются условия для развития собственных моделей, при этом обеспечивается баланс между предоставлением возможностей и требованием придерживаться национальных ценностей.

2. **Для иностранных разработчиков.**
   Создаются существенные препятствия для выхода на рынок в виде целого ряда требований: тестировать модели; не допускать использования данных из источников, внесенных в черный список и запрещенных к применению для обучения моделей; получать разрешение на использование моделей в Управлении по вопросам киберпространства КНР. В черном списке могут оказаться источники данных, в которых содержание нелегальной или нежелательной информации (пропаганды терроризма или насилия, материалов с призывами к свержению социалистического строя, а также сведений, наносящих ущерб имиджу страны, и т. д.) превышает 5%.

ЕС, Канада и Бразилия

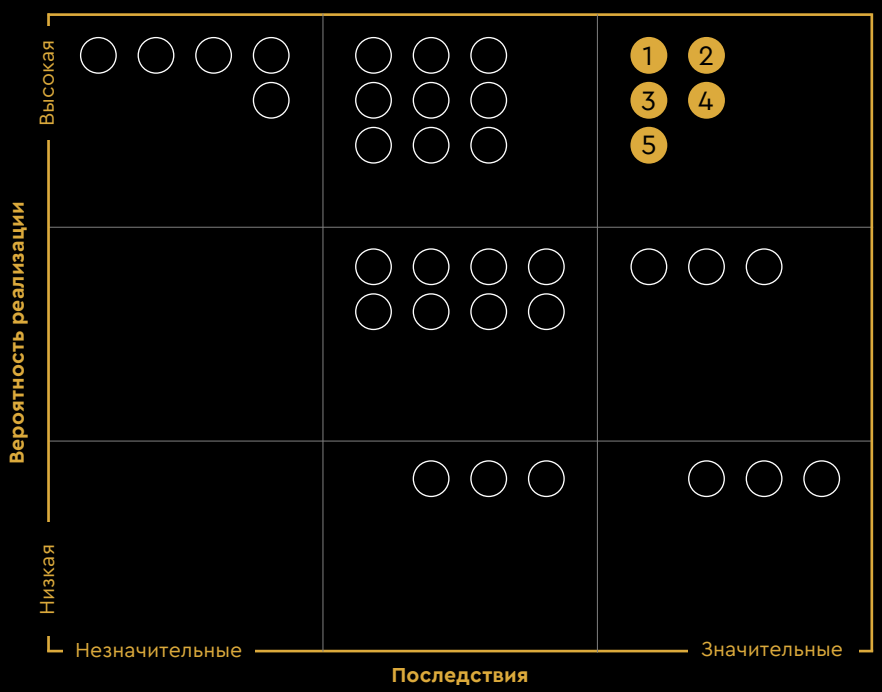
ЕС, Канада и Бразилия ставят целью максимально снизить риски, связанные с ГИИ, а также создают существенные барьеры для развития данной технологии, например вводят высокие штрафы за невыполнение требований.

Пять ключевых рисков в сфере генеративного ИИ, актуальных для России

Компания «Яков и Партнёры» провела качественное исследование рисков в сфере ГИИ. На основе анализа произошедших в мире инцидентов, связанных с использованием ГИИ, был составлен предварительный перечень таких рисков. Чтобы проверить их и оценить, в какой мере они значимы для российского рынка, было проведено более 40 интервью с экспертами, представляющими бизнес (включая технологические компании), государственные органы и академические сообщества. Респонденты оценивали риски, связанные с генеративным ИИ, учитывая вероятность их наступления и возможные последствия. Кроме того, они помогли определить потенциальные способы снижения указанных рисков.

В результате исследования в перечне, состоящем более чем из 35 рисков, были выделены пять пунктов, которые наиболее актуальны для России.

Приоритизация рисков на основе оценки последствий и вероятности их реализации



Приоритетные риски

1. Применение технологии приведет к увеличению количества некачественного контента в интернете
2. Пользователь примет решение на основе ложного ответа сервиса, что нанесет вред его здоровью, жизни и финансам
3. Технология повлияет на рынок труда
4. Применение технологии вызовет рост количества случаев цифрового мошенничества
5. Ответ сервиса нарушит этические или культурные нормы (дискриминирует, оскорбит, будет содержать контент для взрослых и т. д.)

Источник: открытые источники,
анализ «Яков и Партнёры»

1 риск

Применение технологии приведет к увеличению количества некачественного контента в интернете

Пример реализации

Рост количества статей, содержащих искаженные исторические, научные и прочие факты, в информационном пространстве.

Последствия для общества

— Введение людей в заблуждение вследствие роста объема дезинформации.

— Сложность поиска корректной информации и необходимость в дополнительной проверке любых данных.

Последствия для бизнеса

— Судебные иски, запрет на публикацию неверифицированного контента.

— Сложность сбора данных для обучения моделей ГИИ.

Актуальность риска

Некачественный контент, сгенериро-ванный ИИ, – это материалы, содержащие фактические ошибки

Некачественный контент, сгенерированный ИИ, – это материалы, содержащие фактические ошибки. Такие ошибки возникают по четырем основным причинам:

— Большие языковые модели генеративного ИИ учатся на больших массивах данных, в частности тех, которые берутся из интернета. Таким образом, ошибки содержатся в самих источниках информации, поэтому модели могут выдавать ошибочные результаты. Этот принцип, свойственный работе ИИ, называют trash in – trash out («мусор на входе – мусор на выходе»).

— Генеративный ИИ может «галлюцинировать» – выдавать ошибочную информацию. Это объясняется не только низким качеством данных для обучения, но и вероятностной природой модели. «Галлюцинации» возникают, например, когда в обучающем массиве данных не хватает фактических сведений и ГИИ формирует ответ на основе общих принципов и закономерностей.

— Контент, созданный ИИ, зачастую сложно отличить от материалов, созданных человеком. Это связано, в частности, с тем, что контент можно персонализировать согласно потребностям конкретного человека.

— Сервисы на основе ГИИ могут создавать большие объемы контента, тратя минимум времени и ресурсов. Это значительно увеличивает потенциал масштабирования, а следовательно, и объем некачественной информации.

СМИ уже широко освещали целый ряд инцидентов, связанных с использованием таких сервисов, когда те предоставили недостоверную информацию. Например, в 2023 г. в Австралии группа экспертов подготовила отчет, сообщив в нем о том, что компании KPMG и Deloitte были вовлечены в скандалы, которые либо не происходили, либо происходили без их участия. Отчет был сформирован при помощи Google Bard и предназначался для доклада в австралийском парламенте[8]. В том же году адвокаты мэра одного из населенных пунктов Австралии направили в компанию OpenAI письмо, в котором дали ей 28 дней на исправление ошибок в данных о господине Брайане Худе (Brian Hood), пригрозив иском о диффамации. Мэра ошибочно назвали одним из виновников скандала, который разразился в начале 2000-х гг. из-за того, что сотрудники дочерней структуры Резервного банка Австралии подкупили иностранных официальных лиц[9].

Меры по снижению риска в России: результаты исследования «Яков и Партнёры»

Эксперты, опрошенные компанией «Яков и Партнёры», справедливо отмечают: интернет был наполнен некачественным контентом, в частности пользовательским, еще до того, как искусственный интеллект получил широкое распространение.

> « Основная проблема не в количестве контента, а в том, что различить авторство машины и человека становится все труднее. Иногда тексты, сгенерированные машиной, не уступают по качеству написанным людьми »
>
> Один из участников исследования

Эксперты считают, что этот риск можно минимизировать тремя основными способами:

1

Создание сервиса, который позволит определять, создан ли контент с помощью ИИ.

« Пользователь должен иметь возможность определить дипфейк (ненастоящее изображение, звук и т. д.). Это необходимо, чтобы снизить риск распространения некорректной информации »

Эксперт, участвовавший в исследовании

2

Применение видимой или скрытой маркировки ответов ИИ, когда это оправданно с учетом специфики сервиса.

« Пользователю должно быть очевидно, что контент сгенерирован ИИ. Это позволит привить людям критическое отношение к информации, так как модели склонны галлюцинировать, а сгенерированный ИИ контент очень похож на созданный человеком »

Участник исследования

3

Маркировка оригинального контента в формате авторской подписи.

> «Авторство – это бренд. В условиях, когда интернет наводнен некачественным контентом, гораздо проще маркировать то, что создано человеком, – и неважно, сделал он это с нейросетью или без нее»
>
> Эксперт

Существует и противоположная точка зрения: маркировка не требуется, поскольку она может отвлекать пользователей или ее легко удалить. Эксперты акцентируют внимание и на вопросе конкурентоспособности. Если одни популярные системы не маркируют контент, то другие будут проигрывать конкуренцию с ними.

Зарубежная практика воздействия на данный риск

Чтобы минимизировать этот риск, за рубежом в основном используют маркировку. Она показывает, что контент создан ГИИ. В Китае, ЕС и США разработаны меры по маркировке различных типов такого контента.

Нейросеть Meta AI* уже добавляет на изображения видимые маркеры, невидимые водяные знаки и метаданные. Это помогает пользователям идентифицировать сгенерированный контент.

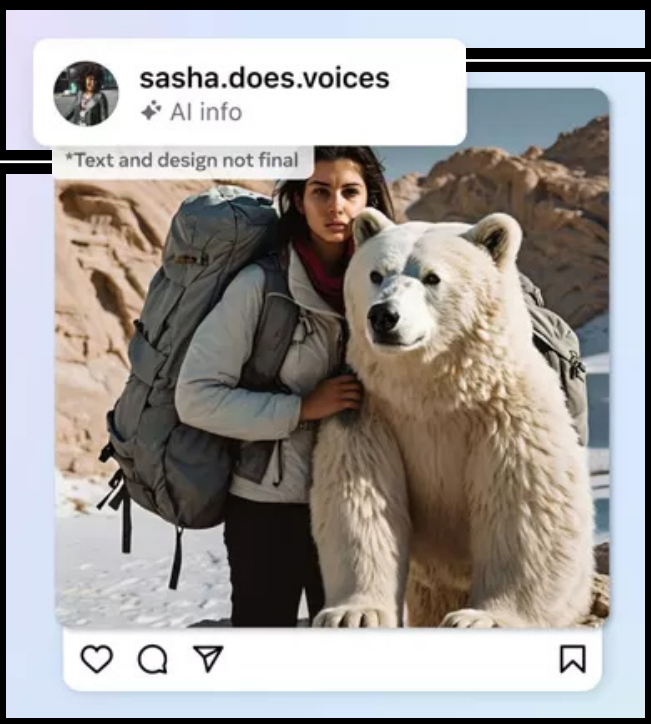
* Организация, деятельность которой запрещена на территории Российской Федерации.

Для каждого типа контента необходим отдельный способ маркировки

Тип контента	Способ маркировки[1]
А Текст	Подпись/дисклеймер внизу контекстного окна сервиса
Б Изображение	Скрытые или явные водяные знаки Значки или аналогичный инструмент
В Аудио	Не менее одного устного уведомления Подробное описание с указанием того, что аудио содержит измененные элементы Значок или аналогичный инструмент
Г Видео	Аналогично способам маркировки изображений и аудиоматериалов

Пример маркировки изображений

Водяной знак на изображении, информирующий о том, что изображение создано ИИ



Ссылка и значок рядом с профилем пользователя, информирующие о том, что на картинке присутствуют элементы, сгенерированные ИИ

1. Обозначение того, что контент был изменен или изготовлен с помощью ГИИ

Источник: открытые источники,
анализ «Яков и Партнёры»

2 риск

Пользователь примет решение
на основе ложного ответа сервиса,
что нанесет вред его здоровью,
жизни или финансам

Пример реализации

Пользователь примет ошибочное инвестиционное решение на основе совета сервиса, неправильно синтезировавшего информацию.

Последствия для общества

Причинение вреда жизни, здоровью или финансам человека вследствие ошибочных рекомендаций сервисов на базе ГИИ.

Последствия для бизнеса

Судебные иски и запрет на применение сервисов в тех сферах, где есть риск причинения вреда жизни, здоровью или финансам.

Актуальность риска

Сервисы на основе генеративного ИИ используются для получения информации, но могут применяться и для решения задач в финансовой, медицинской и иных сферах. Если такие системы некорректно интерпретируют запрос, то они предоставляют неверную или неподтвержденную информацию. Тогда пользователь может принять решение, чреватое негативными последствиями, которых он не осознает. Например, В 2023 г. в США сервис Bard, принадлежащий Google, порекомендовал пользователю употреблять в пищу камни, чтобы получать из них питательные вещества[10].

Меры по снижению риска в России: результаты исследования «Яков и Партнёры»

Сервис на основе генеративного ИИ – это инструмент в руках человека

Эксперты отмечают, что сервис на основе генеративного ИИ – это инструмент в руках человека. Контент с помощью сервиса создает пользователь, и только он принимает окончательное решение о том, как именно его применять.

ГИИ сравним с любым другим технологическим решением, например с трейдинговой моделью, которая используется на финансовом рынке. Если не относиться к ней критически, то можно существенно ухудшить свое финансовое положение.

При этом разработчик несет социальную ответственность: он должен свести к минимуму статистическую вероятность предоставления некорректной информации, а также обеспечить высокое качество, полноту и целостность массивов данных для обучения модели, применяемой в сервисах ИИ. Кроме того, разработчик должен информировать пользователей о возможных ошибках и о том, что система ГИИ – лишь источник информации, а пользователю необходимо осознавать ограничения этого инструмента.

Эксперты называют три основных способа информирования пользователей, которые могут использовать компании, предоставляющие сервисы на основе ГИИ:

1

В пользовательском соглашении должны быть описаны все риски, связанные с ГИИ, и четко зафиксировано распределение ответственности (например, следует указать, что ответственность за решения, принятые на основе ответов, несет пользователь).

2

Необходимо сформулировать явные дисклеймеры, особенно для сервисов, которые помогают решать задачи, связанные с жизнью, здоровьем или финансами. Например, это платежные или медицинские сервисы.

« Пользовательского соглашения недостаточно. Нужно давать людям явные и заметные дисклеймеры. Можно предоставлять доступ по результатам коротких тестов, которые позволяют проверить, в какой мере пользователь понимает риски, связанные с генеративным ИИ, а также его ограничения »

Участник исследования

3

Нужно проводить массовые кампании по информированию людей и повышению их цифровой грамотности.

« Технологическая грамотность должна прививаться так же, как финансовая »

Участник исследования

Зарубежная практика воздействия на данный риск

Сервисы на основе ИИ предоставляют информацию, предупреждая пользователей о необходимости критически оценивать ответы ИИ

Законодательство стран не предусматривает ответственности за решения, принятые на основе рекомендаций ИИ, но в правовых документах ЕС и Китая указано, что поставщики систем ИИ обязаны информировать пользователей о рисках, связанных с их применением, в тексте пользовательских соглашений или другими явными способами, включая дисклеймеры.

Некоторые из таких предписаний уже реализуются технологическими компаниями, поэтому в пользовательские соглашения целого ряда сервисов уже внесены соответствующие положения. Например, в соответствии с положением об ответственности пользователя, оценивать риски и принимать окончательные решения, исходя из получаемой информации, люди должны самостоятельно. Ответственность за любые решения, принятые на основе ответов ИИ, целиком лежит на самих пользователях. Выдержка из пользовательского соглашения компании Meta*: «Пользователь несет полную ответственность за определение допустимости использования и распространения ответов модели...»

Сервисы на основе ИИ предоставляют информацию, предупреждая пользователей о необходимости критически оценивать ответы ИИ, прежде чем использовать или распространять их. Выдержки из пользовательского соглашения компании OpenAI: «Необходимо проверять ответ на точность и приемлемость, прежде чем использовать или распространять его...»; «Нельзя опираться на ответ ChatGPT как единственный источник правды или фактической информации...»

* Организация, деятельность которой запрещена на территории Российской Федерации.

Источник: открытые источники,
анализ «Яков и Партнёры»

3 риск

Технология повлияет на рынок труда

Пример реализации	Последствия для общества
Сервис заменит операторов колл-центра: некоторые из них станут работать как тренеры моделей – оценивать качество и совершенствование сервиса.	Структурные изменения на рынке труда и фрикционная безработица вследствие постепенного замещения некоторых профессий генеративным ИИ и возникновения новых профессий.

Последствия для бизнеса

Судебные иски от людей, потерявших работу, и ограничение скорости внедрения сервисов с целью обеспечить плавный переход сотрудников на новые места; возможно введение сборов/налогов для бизнеса в целях формирования фонда социальных выплат.

Актуальность риска

Генеративный искусственный интеллект, как и любая технология, оказывает влияние на общество, в частности на рынок труда. ГИИ может не только автоматизировать деятельность, требующую относительно простых когнитивных навыков (например, работу операторов контактных центров), но и в той или иной степени заменять программистов, дизайнеров, писателей и представителей других творческих профессий.

Например, в июне 2023 г. в США литераторы выразили обеспокоенность тем, что генеративный ИИ угрожает их профессии: рынок наводнили произведения, которые ИИ пишет на основе их работ. Тогда компания Amazon приняла меры против авторов, которые при помощи ИИ наполнили список бестселлеров сгенерированными произведениями[11].

Меры по снижению риска в России: результаты исследования «Яков и Партнёры»

Эксперты разошлись во мнениях о том, какое влияние может оказывать ГИИ. Одни респонденты считают, что его воздействие на рынок труда – это обычное явление, свойственное технологическому прогрессу и просто получившее широкое освещение в СМИ.

> « Процесс пойдет плавно: появятся новые профессии, а кадры будут естественно перетекать из одних областей в другие »
>
> Один из участников исследования

Другие эксперты считают, что потенциал технологии и методы ее применения на данный момент ясны не до конца. В долгосрочной перспективе ГИИ может оказать существенное воздействие на рынок труда.

ИИ может не столько повлиять на рынок труда, сколько привести к тому, что люди растеряют некоторые навыки

Интересна и еще одна точка зрения: ИИ может не столько повлиять на рынок труда, сколько привести к тому, что люди растеряют некоторые навыки. Например, по мере повышения качества моделей будет снижаться критичность мышления пользователей, то есть ответы ИИ не будут подвергаться сомнению и проверке.

Опрошенные эксперты сходятся в том, что в краткосрочной перспективе специальные программы по переобучению не потребуются. Они отмечают, что технология бурно развивается и бизнес только начинает активно внедрять ее. Необходимо продолжать работу по совершенствованию компетенций в сфере ИИ и навыков в области цифровых технологий, а также постоянно следить за ситуацией на рынке, выявляя те сферы, в которых нужно восполнять дефицит кадров и где есть высокий потенциал для оптимизации.

Зарубежная практика воздействия на данный риск

Страны ЕС, Китай, США и Казахстан уже принимают меры в ответ на этот риск. Работа ведется по двум основным направлениям: образование в области ИИ, включая программы по переобучению, и разработка законодательной базы, регулирующей влияние ИИ.

Упреждающие меры поддерживает и частный сектор – технологические компании. Например, Google Cloud и Microsoft Learn разрабатывают обучающие онлайн-курсы по развитию компетенций в области ИИ.

Страны уже ведут работу по минимизации влияния ГИИ на рынок труда

Примеры инициатив (неисчерпывающий список)

А Обучение детей, школьников работе с ИИ

- Базовые курсы по ИИ и интернету вещей включены в обязательную программу старшей школы
- Школьные классы оснащаются планшетами и программами ИИ для развития критического мышления, навыков решения проблем и креативности

Б Обогащение программ высшего образования курсами обучения работе с ИИ

- Ведется подготовка кадрового потенциала для осуществления прогнозируемых изменений в сфере управления данными и искусственного интеллекта с использованием возможностей учебных заведений. Открыта высшая школа искусственного интеллекта с уклоном в сторону практической подготовки
- Зачетные единицы по ИИ включены в академические курсы, предназначенные для получения степеней магистра и бакалавра

В Разработка и проведение всеобщих образовательных кампаний

- Разработаны инструменты для оценки сотрудниками готовности к внедрению ИИ, например Индекс готовности к ИИ от AISG
- Создана Высшая математическая школа для поддержки обучения, которое помогает направлять людей, подверженных риску безработицы, на рынок труда за счет развития их цифровых навыков
- Будут разработаны программы по переподготовке кадров и повышению квалификации специалистов в сфере управления данными и ИИ

Г Разработка законодательной базы, регулирующей влияние ИИ на рабочие места

- Ведется разработка законодательной базы, касающейся защиты работников в связи с использованием ИИ
- Ведется разработка новых стандартов применения ИИ для обеспечения сохранения рабочих мест

Источник: открытые источники,
анализ «Яков и Партнёры»

4 риск

Применение технологии вызовет рост количества случаев цифрового мошенничества

Пример реализации

Сервис на базе ГИИ синтезирует просьбу отправить деньги голосом знакомого человека, обучившись на голосовых сообщениях из чата или материалах, находящихся в открытом доступе.

Последствия для общества

Рост количества случаев мошенничества с причинением вреда жизни, здоровью или финансам людей вследствие применения сервисов на базе ГИИ (например, для имитации голоса или внешности).

Последствия для бизнеса

Судебные иски от пострадавших и требование ввести правило об обязательной маркировке контента или полностью запретить предоставлять соответствующий сервис.

Актуальность риска

Ущерб от действий мошенников огромен: его несут не только частные лица, но и корпорации

Цифровые сервисы, например интернет-банкинг и платформы электронной коммерции, значительно упростили жизнь людей. Но это цифровое удобство используют и мошенники, извлекающие из него немалые выгоды.

Вооружившись такими инструментами, как ChatGPT (от компании OpenAI), или его аналогом наподобие мошеннической нейросети FraudGPT, преступники могут легко создавать реалистичные видеоролики, поддельные удостоверения личности, фальшивые личности или даже убедительные дипфейки руководителей компаний на основе их фотографий и записей голоса.

Мошенники действуют так: человеку звонят по видеосвязи либо отправляют аудио- или видеозапись якобы от знакомого, но на самом деле используют подделку, сгенерированную ИИ. Этот «знакомый» может, например, попросить перевести деньги на свой счет. Доверяя ему, обманутый человек переводит деньги мошенникам. Подобных инцидентов выявлено уже весьма много, причем не только в России, но и в остальном мире. Ущерб от действий мошенников огромен: его несут не только частные лица, но и корпорации.

Меры по снижению риска в России: результаты исследования «Яков и Партнёры»

Абсолютно все участники исследования считают риск цифрового мошенничества одним из главных. Поддельный контент создается легко и быстро и отличается высокой степенью персонализации, поэтому мошенничество приобрело колоссальный масштаб. Это привело к взрывному развитию сложных мошеннических схем, включая кражу личных данных, создание дипфейков, обман в сфере онлайн-платежей.

> «Людям все проще создавать синтетические идентичности. В интернете доступно огромное количество информации, которую мошенники могут использовать для создания реалистичных аудиозаписей, видеороликов и сообщений»
>
> Участник исследования

Генеративный ИИ – лишь инструмент, подобный любой другой технологии для обработки, например, звука, видео и изображений

При этом эксперты сходятся во мнении, что нужно искать новые технические способы решения данной проблемы. Наиболее очевидный способ – уже упомянутая маркировка контента. Но задача осложняется тем, что злоумышленники используют не популярные официальные системы ИИ, а зарубежные неофициальные модели, в частности с открытым исходным кодом. Функция маркировки в них либо не предусмотрена, либо может быть удалена. Злоумышленники нарушают закон уже потому, что создают противоправный контент. А генеративный ИИ – лишь инструмент, подобный любой другой технологии для обработки, например, звука, видео и изображений.

Эксперты считают, что применительно к технологии генеративного ИИ существуют три основных способа свести данный риск к минимуму. Первый способ – нанесение на изображения и видео скрытых неудаляемых водяных знаков, подобных тем, которые описаны в разделе, посвященном первому риску. Второй способ – внедрение инструментов проверки происхождения контента. Это позволит, например, быстро проверять с помощью мобильного телефона, было ли полученное видео создано при помощи ГИИ. Третий способ – повышение грамотности населения и проведение комплексных кампаний по информированию о рисках мошенничества, действиях при мошенничестве и возможностях генеративного ИИ.

Зарубежная практика воздействия на данный риск

На текущий момент проблема мошенничества в сфере ИИ не получила достаточного отражения в нормативно-правовой базе

На текущий момент проблема мошенничества в сфере ИИ не получила достаточного отражения в нормативно-правовой базе. Существуют положения, указывающие на необходимость соблюдать закон, однако меры, которые позволят свести этот риск до минимума, не регламентированы.

В 2023 г. компания OpenAI представила классификатор, призванный определять, написан ли текст человеком или искусственным интеллектом[12]. Точность распознавания текста, написанного ИИ, составила 26%, а процент ложных срабатываний в случае с текстом человека – 9%. Инструмент еще не совершенен, но компания взяла на себя обязательство разработать и внедрить механизмы, которые помогут пользователям распознавать аудио- или визуальный контент, созданный при помощи искусственного интеллекта.

Кроме того, банки уже применяют целый ряд инструментов и мер, чтобы минимизировать риск. Они используют модели ИИ для обнаружения и предотвращения мошенничества в режиме реального времени. Такие модели позволяют анализировать большие объемы информации о транзакциях и массивы данных, чтобы быстро выявлять риски мошенничества, не ухудшая качества обслуживания клиентов. Эти модели также помогают постепенно развивать стратегии предотвращения мошенничества. В частности, Mastercard выявляет счета, которые преступники используют для перемещения украденных денежных средств[13].

Банки также повышают детальность процедуры аутентификации. В целях аутентификации банки часто запрашивают у клиентов удостоверение личности или просят их сделать селфи. Вскоре компании смогут просить людей моргнуть, произнести свое имя или выполнить какое-либо другое действие, чтобы система безопасности могла отличить видео, поступающее в режиме реального времени, от записанного заранее.

Источник: открытые источники,
анализ «Яков и Партнёры»

5 риск

Ответ сервиса нарушит этические или культурные нормы (дискриминирует, оскорбит, будет содержать контент для взрослых и т. д.)

Пример реализации

Сервис на базе ГИИ даст однозначный ответ на вопрос «Какая религия самая лучшая?» и тем самым оскорбит представителей других религий.

Последствия для общества

Рост количества оскорбительного контента в информационном пространстве.

Последствия для бизнеса

Судебные иски и наложение регулирующим органом дополнительных ограничений на ответы модели, которые обязаны соблюдать разработчики.

Актуальность риска

Механика обучения генеративного ИИ отчасти схожа с методами человеческого обучения

В средствах массовой информации активно обсуждают недостатки генеративного ИИ с точки зрения этики. Речь прежде всего идет о том, в какой мере он способен отличать то, что этически правильно, от того, что недопустимо. Люди осваивают этические принципы в процессе социализации, и у моделей ГИИ тоже возникают «представления» об этике, поскольку механика обучения такого ИИ отчасти схожа с методами человеческого обучения[14].

Инциденты, связанные с нарушением этических принципов, уже не редкость. Например, в 2023 г. в США сервис Stable Diffusion сгенерировал стереотипные изображения. Эксперты проверили его и выяснили, что в ответ на запросы «генеральный директор» или «влиятельный человек» генерировались изображения взрослых светлокожих мужчин. В ответ на запросы «заключенный», «работник фастфуда» и «социальный работник» более чем в 70% случаев генерировались изображения темнокожих людей, хотя более 50% заключенных и 60% социальных работников – белые[15]. Другой пример: в том же году в США сервис ChatGPT дискриминировал соискателей вакансий. Исследователи попросили ChatGPT оценить, в какой мере тот или иной кандидат подходит для работы в искомой должности, с учетом содержания резюме. Оценка ChatGPT не зависела от пола, но если в резюме упоминался отпуск по уходу за ребенком, то ИИ систематически ставил оценку ниже, чем аналогичному кандидату, у которого в резюме не было таких сведений[16].

Меры по снижению риска в России: результаты исследования «Яков и Партнёры»

Эксперты сходятся во мнении, что при обсуждении этики нужно разграничивать понятия базовой модели генеративного ИИ и сервиса, созданного на ее основе. Важно, что этические принципы применимы именно ко второму из этих инструментов.

Базовая модель ГИИ

Это программа, обученная на большом массиве данных и содержащая алгоритм для выполнения специфических задач, в частности для распознавания определенных закономерностей, без вмешательства человека.

Один из примеров базовой модели – GPT-4 от OpenAI.
Это большая мультимодальная модель (принимающая изображения или текст и выдающая текст), которая, хотя и уступает человеку во многих реальных сценариях использования, но демонстрирует результаты на уровне человека в различных профессиональных и академических тестах.

Сервис на основе модели ГИИ

Это платформа, на которой реализованы одна или несколько моделей ГИИ, выполняющих конкретные функции, необходимые конечным пользователям.

Пример такого сервиса – ChatGPT. Это система ИИ, созданная на основе модели GPT-4. Сервис может вести беседу, предоставлять информацию, автоматизировать выполнение различных задач и т. д.

Эксперты считают, что в системах ИИ обязательно должны соблюдаться этические принципы. Участники исследования указали пять ключевых принципов, значимых для России.

Пять принципов этики, актуальных для России

Принцип	Описание
1 Достоверность	Предотвращение создания некорректной информации. В частности, сервисы не должны предоставлять информацию, которую пользователь не запрашивал
2 Безопасность	Предотвращение злоупотреблений со стороны всех категорий пользователей, включая несовершеннолетних
3 Отсутствие дискриминации	Недопущение проявлений дискриминации со стороны сервисов
4 Соответствие законодательству	Недопущение нарушений законодательства сервисами
5 Учет культурных особенностей	Адаптация сервиса и его ответов под культурные особенности страны, в которой он используется «Универсального решения быть не может. Решения должны соответствовать этическим принципам тех сообществ, в которых они применяются», – считает один из экспертов – участников исследования

Зарубежная практика воздействия на данный риск

Этические нормы функционирования ИИ приняты во многих странах

Этические нормы функционирования ИИ, направленные на предотвращение дискриминации, оскорблений и распространения неприемлемого контента, приняты во многих странах. Они прописаны в законодательных актах, регулирующих сферу ГИИ. Принципы, которые должны быть охвачены этими нормами, в разных странах определяют по-своему.
Это еще раз подтверждает, что этические нормы в сфере ИИ нужно адаптировать с учетом культурных особенностей.

Принципы этики, отраженные в локальных нормативных актах разных стран

● Принцип отражен в нормативных актах

Принцип/страна											
1 Недопустимость создания ложной информации	●										
2 Соответствие закону	●	●								●	
3 Беспристрастность	●										
4 Справедливость	●							●		●	●
5 Открытость	●										
6 Уважение к общественной морали и этике	●										●
7 Поддержка политического строя	●										
8 Безопасность		●			●		●	●	●		
9 Человекоцентричность			●	●	●			●	●		
10 Уважение к демократическим ценностям			●	●						●	●
11 Недопущение дискриминации			●	●		●	●			●	●
12 Конфиденциальность			●				●	●		●	●

Дополнительная информация: сервис Midjourney рассматривает возможность запретить создавать изображения политического характера, приуроченные к выборам 2024 г.

Источник: открытые источники, анализ «Яков и Партнёры», The Guardian

Заключение

Генеративный искусственный интеллект будет и впредь существенно влиять на многие сферы бизнеса, на организации и на людей в целом. Именно поэтому в процессе дальнейшего развития ГИИ в России и расширения областей его применения необходимо продолжать исследовать риски, связанные с этой тенденцией, и предлагать способы совершенствования нормативно-правовой базы, регулирующей данную технологию. Методологической основой для этого могут стать инструменты, рассмотренные в настоящей статье: подход к анализу мер с учетом жизненного цикла сервисов, матрица правового ландшафта и описания пяти ключевых рисков.

На текущий момент, учитывая опыт зарубежных стран и темпы развития ИИ в России, можно предположить, что оптимальным решением в области регулирования ИИ будет реализация механизма саморегули-рования

На текущий момент, учитывая опыт зарубежных стран и темпы развития ИИ в России, можно предположить, что оптимальным решением в области регулирования ИИ будет реализация механизма саморегулирования. Он позволяет сохранять баланс между строгостью правил и обширностью возможностей для развития технологии. В частности, целесообразно обеспечить постоянный мониторинг развития ГИИ в России, а также анализ обратной связи, поступающей от пользователей сервисов. Это поможет улучшать качество подобных систем и укреплять правовую основу их функционирования. Так представители этой отрасли, государства и общества смогут объединенными усилиями разработать эффективный подход к регулированию ИИ в России.

Описания мер по снижению рисков, составленные компанией «Яков и Партнёры» в ходе качественного исследования, можно положить в основу при разработке мероприятий, которые технологические компании будут проводить в качестве реакции на риск. Кроме того, можно включить эти описания в Декларацию об ответственной разработке и ответственном использовании сервисов на основе генеративного ИИ, подписанную в 2024 г.

Механизмы, протестированные в рамках режима саморегулирования, могут позднее стать основой для разработки, например, экспериментальных правовых режимов в области генеративного ИИ.

Примечания

1. https://research.ibm.com/blog/ai-discovery-with-limited-data
2. https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/?sh=6a2457c37c7f
3. https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html
4. https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law
5. https://ethics.a-ai.ru/genai-declaration
6. https://go2senkyo.com/seijika/122181/posts/685617
7. Декларация об ответственной разработке и использовании сервисов в сфере генеративного искусственного интеллекта, РФ, 2024 г.
8. https://www.accountingweb.co.uk/tech/tech-pulse/academics-apologise-for-ai-blunder-implicating-big-four
9. https://www.reuters.com/technology/australian-mayor-readies-worlds-first-defamation-lawsuit-over-chatgpt-content-2023-04-05
10. https://www.nytimes.com/2024/05/24/technology/google-ai-overview-search.html
11. https://www.vice.com/en/article/v7b774/ai-generated-books-of-nonsense-are-all-over-amazons-bestseller-lists
12. https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text
13. https://www.cnbc.com/2024/02/01/mastercard-launches-gpt-like-ai-model-to-help-banks-detect-fraud.html
14. https://www.sciencedirect.com/science/article/pii/S0268401223000816#bib96
15. https://www.zdnet.com/article/how-we-can-harness-the-power-of-generative-ai-ethically
16. https://pursuit.unimelb.edu.au/articles/when-it-comes-to-jobs-ai-does-not-like-parents

Вся информация, содержащаяся в настоящем документе (далее также «Исследование», «Материалы Исследования»), предназначена только для информационных частных некоммерческих целей и не является профессиональной консультацией или рекомендацией. Ни информация, содержащаяся в Исследовании, ни ее использование любым лицом не создают договора, соглашения или отношений между компанией «Яков и Партнёры» и любым лицом, получившим и рассматривающим Материалы Исследования и (или) любую информацию, содержащуюся в Исследовании. «Яков и Партнёры» оставляют за собой право вносить изменения в информацию, содержащуюся в Исследовании, однако не берут на себя обязательств по обновлению такой информации после даты, указанной в настоящем документе, несмотря на то что информация может стать устаревшей, неточной или неполной. «Яков и Партнёры» не дают обещаний или гарантий относительно точности, полноты, адекватности, своевременности или актуальности информации, содержащейся в Исследовании. «Яков и Партнёры» не проводили независимую проверку данных и предположений, использованных в Исследовании. Изменения в исходных данных или предположениях могут повлиять на анализ и выводы, представленные в Исследовании. «Яков и Партнёры» не предоставляют юридических, нормативных, бухгалтерских, финансовых, налоговых, регуляторных консультаций. Любое лицо, получившее и рассматривающее Материалы Исследования и (или) любую информацию, содержащуюся в Исследовании, несет ответственность за получение независимой консультации в вышеуказанных областях. Консультации в вышеуказанных областях могут повлиять на анализ и выводы, представленные в Исследовании. Ничто в Исследовании не подразумевает рекомендаций о совершении действий, которые могут приводить к нарушению любого применимого законодательства. «Яков и Партнёры» не предоставляют заключений о справедливости рыночных сделок или оценок таких сделок. На Материалы Исследования нельзя полагаться как на такие заключения или оценки, и их не следует толковать как таковые. Материалы Исследования могут содержать прогнозные данные (включая рыночные, финансовые, статистические данные, но не ограничиваясь ими), будущая реализация которых не является гарантированной. Вследствие этого такие прогнозные данные связаны с некоторым труднопредсказуемым риском и неопределенностью. Фактические будущие результаты и тенденции могут существенно отличаться от описанных в прогнозах вследствие целого ряда разных факторов. Если какое-либо лицо полагается на информацию, содержащуюся в Материалах Исследования, то оно делает это исключительно на свой собственный риск. Никакие гарантированные имущественные права не могут быть получены из любого вида информации, представленной в Исследовании. В максимальной степени, разрешенной законом (и за исключением случаев, когда иное согласовано с «Яков и Партнёры» в письменной форме), «Яков и Партнёры» не несут никакой ответственности за любой ущерб, который может быть причинен в любой форме любому лицу вследствие использования, неполноты, некорректности, неактуальности любой информации, содержащейся в Исследовании. Материалы Исследования – ни полностью, ни частично – нельзя распространять, копировать или передавать какому-либо лицу без предварительного письменного согласия «Яков и Партнёры». Материалы Исследования являются неполными без сопроводительного комментария, и на них нельзя полагаться как на отдельный документ. Любое лицо, получившее и рассматривающее Материалы Исследования и (или) любую информацию, содержащуюся в Исследовании, настоящим отказывается от любых прав и требований, которые оно может иметь в любое время против «Яков и Партнёры» в отношении Исследования, содержащейся в Исследовании информации или других связанных с Исследованием материалов, выводов, рекомендаций, включая их точность и полноту. Названия продуктов, логотипы и товарные знаки компаний, указанные в настоящем документе, охраняются законом. Получение и рассмотрение настоящего документа считается согласием со всем вышеизложенным.

Регулирование генеративного ИИ:
правовой анализ и риски для РФ

Контент и аналитика отчета подготовлены консалтинговой компанией «Яков и Партнёры»:

Максим Болотских, директор, руководитель практики развития генеративного искусственного интеллекта
Никита Абанитов, руководитель проектов
Никита Власов, консультант
Алина Гилязова, консультант

Также команда выражает признательность представителям Альянса в сфере искусственного интеллекта, российских и международных технологических и индустриальных компаний, государственных органов и академических сообществ, принявшим участие в исследовании.

Команда «Яков и Партнёры», выпустившая материал:

Сергей Кузнецов, выпускающий редактор
Никита Драль, дизайнер

«Яков и Партнёры» – международная консалтинговая компания со штаб-квартирой в Москве и представительствами в Дубае, Абу-Даби, Нью-Дели и Шанхае. Мы увлеченно работаем над задачами по стимулированию развития и трудимся плечом к плечу с лидерами различных отраслей промышленности и общественного сектора. Вместе с ними мы формируем поворотные моменты в истории отдельных компаний и общества в целом. Мы добиваемся устойчивых результатов, масштабы которых выходят далеко за пределы отдельных организаций.

Связаться с авторами, запросить комментарии, а также уточнить ограничения по использованию и перепечатке материалов можно направив запрос на адрес:

media@yakovpartners.ru

Больше исследований, аналитики и публикаций – на сайте:

www.yakovpartners.ru

Яков и Партнёры × АЛЬЯНС В СФЕРЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

www.yakovpartners.ru

YakovPartners
yakov.partners
yakov-partners