УДК 004.82

ЭТИЧЕСКИЕ ВЫЗОВЫ ГЕНЕРАТИВНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Вершинин М.В. – магистрант,
Хаймин А.В. – магистрант,
Кузбасский государственный технический университет имени Т.Ф. Горбачева
Россия, г. Кемерово

Развитие генеративного искусственного интеллекта (Generative AI, GAI) в последние годы стало одной из ключевых тенденций цифровой эпохи. Системы, подобные ChatGPT, DALL•E или Midjourney, продемонстрировали способность создавать тексты, изображения и даже аудиовизуальные материалы, неотличимые от результатов человеческого творчества. За 2024–2025 годы GAI эволюционировал в направлении мультимодальности, персонализации и интеграции с реальным временем, что породило новые возможности и риски. Однако столь стремительное распространение подобных технологий породило сложный комплекс этических и социальных проблем. Как отмечают исследователи (MDPI, 2024) и последующие отчёты 2025 года, использование генеративного ИИ требует не только технологической, но и нормативно-ценностной оценки. Цель данной работы – рассмотреть основные риски, связанные с функционированием систем GAI, и обозначить возможные пути их этически устойчивого развития.

Одним из наиболее острых вызовов является вопрос конфиденциальности информации, используемой для обучения нейросетей. Алгоритмы обучаются на огромных массивах данных, часто включающих персональные публикации пользователей сети. При этом сбор подобных данных нередко осуществляется без их согласия, что противоречит базовым принципам защиты личной информации (MDPI, 2024). Кроме того, модели ИИ способны непреднамеренно воспроизводить элементы исходных данных, что создаёт угрозу утечки приватных сведений. В 2024–2025 годах появились новые технические и правовые практики. Технологии вроде дифференциальной приватности и федеративного обучения стали применяться чаще; появились способы представлять данные так, чтобы модель не могла легко восстановить личную информацию (privacy-preserving embeddings), и методы «unlearning» – удаление конкретных примеров из уже обученной модели [1]. Параллельно суды и регуляторы всё активнее требуют отчётности о том, откуда взяты данные и на каком основании их используют. Все это привело к широкому внедрению практик отслеживания происхождения данных (data provenance) и обязательной фиксации метаданных о лицензиях и согласиях. В связи с этим авторы статьи подчёркивают необходимость внедрения принципа data minimization – ограничения объёма используемых данных до минимально необходимого, а также внедрения механизмов их удаления по запросу субъектов; дополнительно сейчас

рекомендуется использовать технологии, которые защищают приватность (например, дифференциальную приватность), методы, гарантирующие фактическое удаление данных из модели, и по возможности хранить чувствительную информацию локально на устройстве пользователя

Проблема алгоритмической предвзятости тесно связана с тем, что ИИ воспроизводит стереотипы, присутствующие в исходных обучающих выборках. Если данные содержат скрытую дискриминацию, система неизбежно её унаследует. Примеры подобных случаев достаточно известны. Так, система COMPAS, применявшаяся в США для оценки вероятности рецидивов преступлений, систематически завышала показатели риска для афроамериканцев. Алгоритм подбора персонала от Amazon демонстрировал гендерную предвзятость, снижая шансы женщин при отборе кандидатов на технические должности. В 2024–2025 годах внимание сместилось на более чёткие и практичные способы проверки справедливости моделей. Появились простые метрики, которые помогают находить перекрёстную дискриминацию (когда люди страдают по нескольким признакам одновременно), стресс-тесты, проверяющие модели на политические и культурные искажения, и инструменты для постоянного мониторинга работы моделей в реальных условиях. Исследователи (MDPI, 2024) делают вывод, что этические фильтры должны быть встроены на ранних этапах проектирования, а не вводиться постфактум. Необходим многоуровневый аудит данных и постоянный мониторинг справедливости решений.

Одним из наиболее обсуждаемых последствий генеративных технологий стало появление феномена deepfake – синтетически созданных изображений и видеозаписей, визуально неотличимых от реальных [3]. Эти технологии способны подрывать доверие к СМИ, искажать общественное мнение и использоваться для манипуляций в политической сфере. По мнению авторов статьи, deepfake можно рассматривать как форму «когнитивного загрязнения» (cognitive pollution), при которой подлинность информации становится неотличимой от симуляции. Для предотвращения подобных рисков предлагается параллельное развитие систем автоматической детекции подделок, а также повышение медиаграмотности населения и формирование критического мышления при восприятии цифрового контента.

Современные модели ИИ отличаются высокой сложностью внутренней архитектуры, что делает невозможным полное объяснение логики их решений. Это явление получило название black box problem. Отсутствие прозрачности создаёт трудности в определении ответственности за ошибки и предвзятые результаты. В контексте медицины, судопроизводства или государственного управления подобная непрозрачность может иметь серьёзные последствия. В качестве решения исследователи предлагают концепцию Explainable AI (XAI) – объяснимого ИИ, где результаты сопровождаются интерпретацией логики вывода [2]. Такая прозрачность позволяет восстановить цепочку рассуждений модели и повысить доверие со стороны пользователей и регуляторов.

Понимая потенциальные риски, государства и международные организации предпринимают шаги к созданию правовых рамок для использования ИИ.

Так, Европейский AI Act (2024) устанавливает классификацию систем по уровням риска и запрещает опасные формы применения, включая социальный рейтинг и скрытую манипуляцию поведением граждан. Рекомендации UNESCO и OECD формулируют общие принципы этики ИИ: прозрачность, подотчётность, справедливость и уважение человеческого достоинства. Крупные компании, в частности OpenAI, также внедряют внутренние кодексы этического регулирования, ограничивая применение своих моделей для вредоносных целей.

В работе подчёркивается, что решение этических проблем невозможно только с помощью технических средств. Необходим междисциплинарный подход, объединяющий инженеров, юристов, философов и представителей общественных наук. Авторы выделяют несколько ключевых направлений:

1) Ethics by Design – проектирование систем с учётом этических принципов с ранних стадий разработки;

2) трассируемость данных – фиксация источников и изменений информации, использованной для обучения моделей;

3) независимый аудит алгоритмов на предмет предвзятости и безопасности;

4) образовательные инициативы, направленные на повышение цифровой и этической грамотности пользователей.

Эти меры позволяют не ограничивать развитие технологий, а формировать условия для их устойчивого и ответственного применения.

Генеративный искусственный интеллект представляет собой двойственную силу: он способен приносить значительную пользу, но одновременно несёт риски, которые нельзя игнорировать [1]. Как справедливо отмечается в статье (MDPI, 2024), этика не должна рассматриваться как тормоз инноваций – напротив, она является условием их устойчивости и социальной приемлемости. Будущее ИИ зависит не только от технических возможностей, но и от того, насколько человечество сумеет встроить в эти системы собственные моральные принципы, обеспечив прозрачность, справедливость и уважение к человеческому достоинству.

Список литературы:

1. Ethics of generative AI and manipulation: a design-oriented research agenda // Springer Link. – URL: https://link.springer.com/article/10.1007/s10676-024-09745-x/ (дата обращения: 16.11.2025).

2. Generative AI, Research Ethics, and Higher Education Research: Insights from a Scientometric Analysis // ResearchGate. – URL: https://www.researchgate.net/publication/381116874_Generative_AI_Research_Ethics_and_Higher_Education_Research_Insights_from_a_Scientometric_Analysis/ (дата обращения: 17.11.2025).

3. Understanding the Ethics of Generative AI: Established and New Ethical Principles // ResearchGate. – URL:

https://www.researchgate.net/publication/387854870_Understanding_the_Ethics_of_Generative_AI_Established_and_New_Ethical_Principles/ (дата обращения: 17.11.2025).