Publishing House "ANALITIKA RODIS" (analitikarodis@yandex.ru) http://publishing-vak.ru/

УДК 34

Генеративный искусственный интеллект и DeepFake-технологии как угроза информационной безопасности граждан

Симонян Альберт Арамович

Аспирант,
кафедра информационного права и цифровых технологий,
Саратовская государственная юридическая академия,
410056, Российская Федерация, Саратов, ул. им. Н.Г. Чернышевского, 104;
e-mail: deewer25092000@yandex.ru

Аннотация

В данной статье рассматриваются основные угрозы применения систем генеративного искусственного интеллекта. Проанализировано существующее определение терминов «DeepFake» («дипфейк»), «синтетический контент» как в рамках российского законодательства, так и в работах ученых. Представлен авторский вариант определения термина «дипфейк» для закрепления его в законодательстве РФ. Предложены правовые и технические методы противодействия угрозам от применения дипфейк-технологий. Подчеркивается необходимость продолжения исследований в сфере правового регулирования ИИ.

Для цитирования в научных исследованиях

Симонян А.А. Генеративный искусственный интеллект и DeepFake-технологии как угроза информационной безопасности граждан // Вопросы российского и международного права. 2025. Том 15. № 2А. С. 165-172.

Ключевые слова

Дипфейк, искусственный интеллект, синтетический контент.

Введение

Стремительное развитие систем искусственного интеллекта привело к большим успехам в таких сферах жизни как бизнес, маркетинг, коммуникация, образование. Однако эти системы могут использоваться и как инструменты для создания различных угроз безопасности граждан. Отдельным видом систем искусственного интеллекта выделяют генеративные системы ИИ, способные синтезировать фотографии, видео и аудио, неотличимые от реальных. Такие системы с невероятной точностью способны создавать лица как существующих людей, например политиков или актеров, так и людей, которых не существует в мире. Также большое развитие получили системы синтеза речи, способные воссоздать голос и интонацию любого человека, опираясь только на несколько минут существующих аудиозаписей голоса этого человека. Данные технологии и продукты их применения называются DeepFake или «дипфейк». Так ярким примером применения таких технологий стало воссоздание голоса умершего родственника. Пока вопрос этичности использования таких технологий остается открытым, однако их применение уже доступно каждому пользователю. Опасность такой технологии заключается в психологическом давлении на различные группы лиц с целью манипуляции общественным мнением граждан, шантажа, мошенничества или клеветы.

Основное содержание

На данный момент известно множество угроз от применения DeepFake технологий. Основными являются:

– Распространение дезинформации и манипуляция общественным мнением. Создание фальшивых новостей, видеообращений политиков или публичных лиц, которые могут провоцировать социальные волнения, политические кризисы или международные конфликты. Например, поддельное видео с экс-президентом США Бараком Обамой, выпущенное в 2018 году, в котором экс-президент высказывается про президента Дональда Трампа. Данное видео было сделано для демонстрации продвинутости DeepFake технологий и призывало общество обратить внимание на такого вида угрозы. Подобные видео используются для предвыборных кампаний с целью подрыва доверия населения к кандидатам. Также известны случаи, когда жертвами дипфейков становились звезды шоу-бизнеса, что является серьезной угрозой их репутации, и что позволяет манипулировать мнением их фанатов.

– Финансовые мошенничества. Сейчас активно развивается применение дипфейков мошенниками, когда с аккаунта руководителя компании в мессенджерах сотрудникам присылают аудиозаписи с голосом руководителя с просьбой о переводе денег. Подобные случаи происходят и с аккаунтов родственников граждан. Также мошенники используют звонки по телефону, имитируя голоса и видео близких людей, коллег и знакомых.

– Нарушение приватности граждан. Огромной проблемой последних лет стало распространение материалов порнографического характера в виде дипфейков существующих людей, как обычных граждан, так и знаменитостей. На основе обычных фотографий из соцсетей чат-боты, оснащенные генеративным искусственным интеллектом, способны создавать синтетический контент порнографического характера, что является нарушением прав граждан на неприкосновенность частной жизни. Также это влечет и серьезные репутационные угрозы.

– Подрыв доверия к цифровым доказательствам. Дипфейки могут усложнять судебные процессы, так как видео- и аудиоматериалы теряют доказательную силу. Сюда включены как фальсификация видео- и аудиоматериалов с места преступления, так и сложность подтверждения, что такие доказательства не являются дипфейками. Важность этой проблемы подчеркивает Лаптев В.А. в своей работе [Лаптев, 2021].

Это лишь неполный список всех известных угроз и вызовов, с которыми сталкивается человечество из-за применения DeepFake технологий.

Важным ограничением в регулировании таких технологий является отсутствие в Российской Федерации законодательства в сфере искусственного интеллекта. Анализируя состояние российского законодательства, можно выделить следующие аспекты. Во-первых, отсутствуют определения понятий «дипфейк» и «синтетический контент», что сильно затрудняет разработку законопроектов в данной сфере. Хочется отметить, что несмотря на введение Федеральный закон от 24 апреля 2020 г. N 123-ФЗ "О проведении эксперимента по установлению специального регулирования в целях создания необходимых условий для разработки и внедрения технологий искусственного интеллекта в субъекте Российской Федерации - городе федерального значения Москве и внесении изменений в статьи 6 и 10 Федерального закона "О персональных данных" (с изменениями и дополнениями), в данном законе также отсутствует определение понятия «дипфейк», хоть и присутствует определение искусственного интеллекта. Данная проблема не позволяет не только определить виноватого в правонарушении, но и часто не позволяет утверждать о факте преступления с точки зрения закона.

Во-вторых, 207.1, 207.2, 207.3 статьи УК РФ «Публичное распространение заведомо ложной информации…» или 128.1 статьи УК РФ «Клевета» могут применяться к дипфейкам, однако умысел доказать крайне сложно из-за отсутствие какого-либо регулирования дипфейк технологий.

В-третьих, в рамках «Кодекса этики в сфере искусственного интеллекта» можно выделить такие общие принципы для систем ИИ, как маркировка контента, оценка рисков, ограничения в применении ИИ в деструктивных целях. Такие рекомендательные документы вносят важный вклад не только в развитие технологий искусственного интеллекта, но и объединяют крупные компании на разработку регламентов использования ИИ для будущих законодательств.

Рассмотрим, как российские ученые подходят к определению понятия «дипфейк». Добробаба рассматривает дипфейк как технологию изготовления поддельных фото и видео, которая, используя методику компьютерного синтеза изображения, основанную на искусственном интеллекте, переносит черты лица с изображения человека на целевое фото (видеозапись) с высокой степенью правдоподобия [Добробаба, 2022]. В работе автор отмечает, что дипфейк-технолгия создает синтетический контент, который может нарушать права человека, включая неприкосновенность частной жизни, честь и достоинство. Она подчеркивает двойственность технологии: с одной стороны, её позитивное применение в образовании и медиа, с другой — риски криминализации. Автор акцентирует необходимость законодательного регулирования, включая обязательную маркировку дипфейк контента и разработку автоматизированных инструментов для его обнаружения. Однако данное определение в какой-то степени теряет актуальность, так как с развитием технологий форм дипфейка стало гораздо больше (аудио, текст).

Бодров и Лебедева отмечают отсутствие закрепленного определения дипфейка в российском законодательстве и предлагают собственную дефиницию, учитывающую

технические и правовые аспекты [Бодров Н.Ф., Лебедева, 2023]. «Дипфейк — цифровой продукт в виде текста, графики, звука или их сочетания, сгенерированный полностью или частично при помощи нейросетевых технологий с целью введения в заблуждение или преодоления пользователем систем контроля и управления доступом». Ключевой акцент сделан на классификацию дипфейков по типу контента (видео, аудио, текст) и их криминогенный потенциал. Авторы подчеркивают, что существующие научные определения часто сужают понятие, игнорируя, например, звуковые дипфейки или замену не только лиц, но и объектов.

Если рассматривать иностранный опыт в определении и регулировании дипфейков, то можно обратиться к новым правилам по регулированию искусственного интеллекта в Китае [Новые правила в области искусственного интеллекта в Китае, www]. В данном документе дано определение глубокому синтезу как технологиям, использующим генеративные и синтетические алгоритмы, такие как глубокое обучение и виртуальная реальность, для генерации текста, изображений, аудио, видео, виртуальных сцен и другой интернет-информации (technologies that utilize generative and synthetic algorithms, such as deep learning and virtual reality, to generate text, image, audio, video, virtual scenes, and other internet information.). Данное определение шире, чем те, что рассматривались выше. Более того в данном акте отдельно выделены виды генерируемого или модифицируемого контента – текст, голос, музыка, биометрический и не биометрический, а также 3D реконструкция объектов и цифровые модели.

Обратимся также к определению понятия дипфейк в Австрии [Aktionsplan Deepfake Wien, 2022]. «Дипфейки — это полностью поддельные видео, изображения или аудио, в которых людей заставляют говорить что-то или которые совершают действия, которых на самом деле никогда не было. Используя мощные методы искусственного интеллекта, видео можно манипулировать таким образом, что уже невозможно будет определить, по крайней мере невооруженным глазом, являются ли они настоящими или были смонтированы». В данном определение акцент сделан на вредоносном характере воздействия таких технологий, и оно больше сфокусировано на нормативно-правовой сущности дипфейков.

Важно отметить, что сложно подобрать наиболее общее определение, способное полностью описать одним предложением многогранность дипфейков. В качестве основных моментов считаем важным выделить общий характер дипфейков, а не только касательно подмены лиц и голоса людей, так как видео, например, с искусственно синтезированным взрывом здания также может считаться дипфейком. Кроме того, к дипфейкам можно отнести и генерацию искусственной музыки или мультипликационных фильмов, что само по себе может и не нести никаких правонарушений. Также стоит подчеркнуть, что в последние годы стали популярны цифровые аватары, которые могут проявлять активность в социальных сетях, общаться с аудиторией, однако они не являются реальными людьми, а полностью созданы с помощью искусственного интеллекта. Такие аватары могут и не являться угрозой обществу, однако без явных пометок их часто сложно отличить от реальных людей, что может вводить в заблуждение.

Предложим авторское определение: «Дипфейк – вид реалистичного цифрового контента, искусственно синтезированного с помощью технологий искусственного интеллекта, представимый в виде текста, звука, голоса, фото, видео и/или их комбинаций, а также в любых других возможных формах интернет информации, целью которого является подражание реальности и введение в заблуждение». Данное определение подчеркивает следующие основные особенности дипфейков:

– технологическая основа: цифровой контент создан с помощью ИИ

– разнообразие форм цифрового контента: фото, видео, звук и т.д.

– нормативный характер: подражание реальности и введение в заблуждение.

Эти составляющие позволят не только в общем смысле слова описать понятие «дипфейк», учитывая современное развитие технологий, но и позволит оттолкнуться от него с правовой точки зрения. Важным этапом правового регулирования дипфейков станет дальнейшая классификация дипфейк-технологий по целям его применения для определения конкретных правонарушений и введения ответственности за них. Стоит подчеркнуть, что сами по себе дипфейки не несут опасности, и существуют множество абсолютно легальных способов их применения. Однако в рамках законодательства необходимо разработать допустимые цели применения дипфейк-технологий. Важным пунктом в будущем законодательстве может быть классификация дипфейков по виду генерируемого контента и запрет некоторых из них для распространения (например, контент порнографического характера), разделение дипфейков на «разрешенные» и «запрещенные».

Важным пунктом в будущем регулировании дипфейков является разработка технических способов выявления синтетического контента. О важности таких мер пишут многие российские и зарубежные ученые. Однако технические автоматизированные решения на основе искусственного интеллекта, направленные на противодействие угрозам от дипфейков, пока являются малоэффективными. Если рассмотреть методы обучения ИИ для генерации синтетического контента, то такие технологии изначально нацелены на обман как людей, так и систем ИИ, призванных их обнаруживать. Иными словами, DeepFake-технологии разрабатываются с учётом обхода существующих механизмов их выявления. Это позволяет злоумышленникам оперативно модернизировать свои инструменты, в то время как методы противодействия остаются в положении «догоняющих». Однако активно разрабатываются альтернативные методы на основе статистики и частотного анализа [Frank J. et al., 2020]. Несмотря на слабости систем для выявления дипфейков, считаем необходимым не только их разработку и регулярное совершенствование, но и создания на их основе общедоступного бесплатного приложения с понятным интерфейсом. Такое приложение, разработанное на государственном уровне, позволило бы не только гражданам проверить различный контент на дипфейк, но и могло бы быть внедрено в различные интернет-ресурсы, как социальные сети, электронные почты и т.д. Данный сервис мог бы высылать предупреждения пользователям при детектировании подозрительного контента, что позволило бы существенно снизить эффективность интернет-мошенничества. Подобные разработки уже внедряются в систему «Антиплагиат» при проверке текстов.

Также на основе проведенного анализа для регулирования DeepFake технологий и предотвращения преступлений в данной сфере считаем необходимым разработку административных и уголовных мер наказаний для всех участников процесса распространения запрещенных видов дипфейков, а не только для разработчиков. Важно, что основной ущерб обществу наносит не просто сгенерированное изображение или видео, а его массовое распространение в сети интернет без должного контроля. Так, например, возможно рассмотреть необходимость введения обязательств для новостных ресурсы в интернете проверять достоверность публикуемых новостей, что позволит заметно сократить массовое распространение фейков и дезинформации. Проверка должна осуществляться с помощью разработанных государственных автоматизированных систем выявления дипфейков. Конечно, как было указано выше, нет гарантий, что данные системы смогут выявить любой синтетический контент. В таком случае новостной ресурс не несет ответственности за публикацию, если она окажется дипфейком. Однако важно обязать СМИ проводить подобный

контроль за публикуемым контентом и хранить отчетность о проверках. В противном случае накладывать штрафы и другие меры наказания. Даже эти меры позволят замедлить распространение вредоносных дипфейков.

Еще одним важным пунктом предотвращения угроз от дипфейков может стать введение обязательств для операторов ИИ по маркировке дипфейк контента. Важно, чтобы новостные ресурсы, соцсети и любые другие организации добавляли к синтетическим данным уведомление в явной недвусмысленной форме для пользователей. Это позволит лучше проинформировать граждан о применении дипфейк технологий. Важность данного подхода отмечают множество российских ученых, а также подобные меры уже закреплены в правовых актах Китая и ЕС.

Часто опасность от дипфейков кроется в приложениях, разработанных для всеобщего пользования. Такие приложения часто не имеют серьезных форм контроля генерируемого контента, что позволяет гражданам использовать их не только для творчества и развлечений, но и для различных правонарушений. В рамках регулирования деятельности разработчиков систем искусственного интеллекта предлагаем создание национального реестра систем искусственного интеллекта (в частности, для систем генерации дипфейков), где каждый разработчик (включая дистрибьюторов иностранного ПО) обязан будет зарегистрировать используемые в публикуемом приложении дипфейк технологии. Данная регистрация может включать в себя информацию по используемым моделям ИИ, общее описание набора данных, который использовался при обучении, описание ограничений, наложенных на генерируемый контент и др. Без такой регистрации (или лицензирования) системы генерации дипфейков, разработчик не должен иметь право на размещение подобного приложения в общее пользование. Важно разработать механизмы и органы контроля по лицензированию дипфейк технологий. Таким образом любые системы ИИ, не зарегистрированные в данном реестре, будут запрещены к использованию. Подобный опыт уже был введен в рамках законодательства Китая и ЕС.

С технической точки зрения у разработчиков систем генеративного искусственного интеллекта есть еще одна возможность для предотвращения угроз – на любой вид цифрового контента, будь то звук, изображение или текст добавлять невидимые «маски» (невидимые специальные символы в текст, скрытые частоты в звук или шум в изображения). Такие «маски» невозможно определить человеческим глазом. Такой индивидуальный подчерк для каждой системы ИИ будет работать как штрихкод, который позволит с помощью специальной программы легко определить происхождение изображения. Такая технология уже давно разработана и практикуется в исследовательских целях. Последние версии известной ChatGPT уже интегрировали этот подход при генерации текстов. Предлагаем для усиления вышеуказанных мер обязать разработчиков интегрировать данную технологию при генерации дипфейков. Кроме того, такой подход позволит отделять лицензированные системы ИИ от нелицензированных. Данная мера может быть достаточно эффективна, так как большинство злоумышленников не обладают возможностями разработать собственные системы ИИ и пользуются доступными решениями, разработанными большими корпорациями.

Заключение

Таким образом, системы искусственного интеллекта для генерации дипфейков развиваются с невероятной скоростью. Сейчас каждый пользователь интернета уже способен самостоятельно опробовать различные достижения прогресса в этой сфере. Однако для того чтобы добиться органичного развития и применения дипфейк технологий, важно с такими же темпами развивать

и правовое регулирование для этой сферы. Не ограничить развитие технологий, а направить и обезопасить. Так в данной статье были предложены авторское определение термина «дипфейк», а также меры по правовому и технологическому регулированию дипфейк технологий для предотвращения и сокращения угроз информационной безопасности граждан.

Библиография

1. Бодров Н.Ф., Лебедева А.К. Понятие дипфейка (deepfake) в российском праве, его классификация и проблемы правового регулирования // Юридический вестник ДГУ. 2023. Т 48. № 4(68). С. 173-181.
2. Добробаба М.Б. Дипфейки как угроза правам человека // Lex russica. 2022. Т. 75. № 11 (192). С. 112-119.
3. Кодекс этики в сфере искусственного интеллекта. М.: Альянс в сфере ИИ, 2023. URL: https://ethics.a-ai.ru (дата обращения: 13.05.2025).
4. Лаптев В.А. Deepfake и иные продукты искусственного интеллекта на пути развития онлайн-правосудия // Актуальные проблемы российского права. 2021. Т. 16. № 11 (132). С. 180-186.
5. Новые правила в области искусственного интеллекта в Китае // Latham & Watkins. 2023. URL: https://ai.gov.ru/knowledgebase/normativnoe-regulirovanie-ii/2023_novye_pravila_v_oblasti_iskusstvennogo_intellekta_v_kitae_china_s_new_ai_regulations_latham_watkins_/ (дата обращения: 13.05.2025).
6. О проведении эксперимента по установлению специального регулирования в целях создания необходимых условий для разработки и внедрения технологий искусственного интеллекта в субъекте Российской Федерации – городе федерального значения Москве и внесении изменений в статьи 6 и 10 Федерального закона «О персональных данных»: федер. закон от 24 апреля 2020 г. № 123-ФЗ // Собрание законодательства РФ. 2020. № 17. Ст. 2728.
7. Уголовный кодекс Российской Федерации от 13 июня 1996 г. № 63-ФЗ (ред. от 05.04.2024) // Собрание законодательства РФ. 1996. № 25. Ст. 2954.
8. Aktionsplan Deepfake Wien, 2022 2 von 30 III-740 der Beilagen XXVII. GP - Bericht - 02 Hauptdokument.
9. European Union Artificial Intelligence Act. 2024. URL: https://artificialintelligenceact.eu (дата обращения: 13.05.2025).
10. Frank J. et al. Leveraging frequency analysis for deep fake image recognition // International conference on machine learning. PMLR, 2020. С. 3247-3258.

Generative artificial intelligence and DeepFake technologies as a threat to citizens' information security

Al'bert A. Simonyan

Postgraduate Student,
Department of Information Law and Digital Technologies,
Saratov State Law Academy,
410056, 104 N.G. Chernyshevskogo str., Saratov, Russian Federation;
e-mail: Deewer25092000@yandex.ru

Abstract

This article analyzes the main threats posed by the use of generative artificial intelligence systems. The existing definitions of the terms "DeepFake" and "synthetic content" within both Russian legislation and academic research are examined. The author proposes a definition of the term "deepfake" for its incorporation into Russian legislation. Legal and technical countermeasures to address threats arising from deepfake technologies are suggested. The article also emphasizes the need for continued research in the field of AI legal regulation.

For citation

Simonyan A.A. (2025) Generativnyi iskusstvennyi intellekt i DeepFake-tekhnologii kak ugroza informatsionnoi bezopasnosti grazhdan [Generative artificial intelligence and DeepFake technologies as a threat to citizens' information security]. *Voprosy rossiiskogo i mezhdunarodnogo prava* [Matters of Russian and International Law], 15 (2A), pp. 165-172.

Keywords

Deepfake, artificial intelligence, synthetic content.

References

1. Aktionsplan Deepfake Vienna, 2022 2 of 30 III-740 of the XXVII. GP - Bericht - 02 Hauptdokument.
2. Bodrov N.F., Lebedeva A.K. (2023) The concept of deepfake in Russian law, its classification and problems of legal regulation. Legal Bulletin of DSU, 48: 4 (68), pp. 173-181.
3. Code of Ethics in the Sphere of Artificial Intelligence. Moscow: Alliance in the Sphere of AI, 2023. Available at: https://ethics.a-ai.ru [Accessed 13.05.2025].
4. Criminal Code of the Russian Federation of June 13, 1996 No. 63-FZ (as amended on 05.04.2024) (1996). Collected Legislation of the Russian Federation, 25. Art. 2954.
5. Dobrobaba M.B. (2022) Deepfakes as a threat to human rights. Lex russica, 75: 11 (192), pp. 112-119.
6. European Union Artificial Intelligence Act. (2024). Available at: https://artificialintelligenceact.eu [Accessed 13.05.2025].
7. Frank J. et al. (2020) Leveraging frequency analysis for deep fake image recognition. International conference on machine learning. PMLR, pp. 3247-3258.
8. Laptev V.A. D(2021)eepfake and other products of artificial intelligence on the way to the development of online justice. Actual problems of Russian law, 16: 11 (132), pp. 180-186.
9. New rules in the field of artificial intelligence in China. Latham & Watkins. (2023). Available at: https://ai.gov.ru/knowledgebase/normativnoe-regulirovanie-ii/2023_novye_pravila_v_oblasti_iskusstvennogo_intellekta_v_kitae_china_s_new_ai_regulations_latham_watkins_/ [Accessed 13.05.2025].
10. On conducting an experiment to establish special regulation in order to create the necessary conditions for the development and implementation of artificial intelligence technologies in a constituent entity of the Russian Federation - the federal city of Moscow and on amendments to Articles 6 and 10 of the Federal Law "On Personal Data": Federal Law of April 24, 2020 No. 123-FZ (2020). Collected Legislation of the Russian Federation, 17. Art. 2728.