Научная статья
УДК/UDC 343.2/.7, 343.9.
DOI: 10.21779/2224-0241-2023-48-4-173-181

Понятие дипфейка (deepfake) в российском праве, его классификация и проблемы правового регулирования

Н.Ф. Бодров[1], А.К. Лебедева[2]
[1, 2] Университет имени О.Е. Кутафина (МГЮА), г. Москва, Российская Федерация
[1] *bodrovnf@gmail.com*, https://orcid.org/0000-0002-9005-3821
[2] *tonya109@yandex.ru*, https://orcid.org/0009-0004-9344-2103

Аннотация. В настоящее в российском законодательстве отсутствуют, какие-либо нормы, регулирующие распространение и использования дипфейков, в научной литературе имеются совершенно полярные точки зрения на необходимость подобного регулирования. Таким образом, в работе поднимается вопрос о необходимости адаптации существующих механизмов правового регулирования с учётом активного развития технологий искусственного интеллекта. В работе авторами предлагается авторская формулировка понятия «дипфейк», описаны угрозы их применения (как потенциальные, так и реальные), а также предлагается видовая и типовая классификация дипфейков. По результатам исследования делается вывод о необходимости проработки механизмов предупреждения использования и распространения дипфейков, например, закрепив нормативно определение «дипфейка», вводя нормы об обязательной маркировке дипфейков и использовании водяных знаков при их создании.

Ключевые слова: дипфейк, генеративный контент, правовое регулирование, специальные знания, маркировка контента

Для цитирования: 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. DOI: 10.21779/2224-0241-2023-48-4-173-181

Благодарности: Статья подготовлена в рамках госзадания «Российская правовая система в реалиях цифровой трансформации общества и государства: адаптация и перспективы реагирования на современные вызовы и угрозы (FSMW-2023-0006)». Регистрационный номер: 1022040700002-6-5.5.1.

Original article

The concept of deep face in Russian law, its classification and problems of legal regulation

Nikolay F. Bodrov[1], Antonina K. Lebedeva[2]
[1, 2] Moscow State Law University (MSAL), Moscow, Russian Federation
[1] *bodrovnf@gmail.com*, https://orcid.org/0000-0002-9005-3821
[2] *tonya109@yandex.ru*, https://orcid.org/0009-0004-9344-2103

Abstract. Currently, there are no any norms regulating the distribution and use of deepfakes in the Russian legislation, in the scientific literature there are completely polar points of view on the need for such regulation. Thus, the paper raises the question of the need to adapt the existing mechanisms of legal regulation in view of the active development of artificial intelligence technologies. The authors propose the author's formulation of the concept of " deepfake", describe the threats of their use (both potential and real), and also propose species and type classification of deepfakes. According to the results of the study concludes that it is necessary to develop mechanisms to prevent the use and distribution of deepfakes, for example, fixing the normative definition of "deepfake", introducing norms on mandatory labeling of deepfake and the use of watermarks in their creation.

Key words: deepfakes, generative content, legal regulation, forensic science, content labeling

For citation: Bodrov N.F., Lebedeva A.K. The concept of deep face in Russian law, its classification and problems of legal regulation. *Yuridicheskii vestnik DGU = Law Herald of DSU*, 2023, vol. 48, no. 4 (68), pp. 173–181. DOI: 10.21779/2224-0241-2023-48-4-173-181 (In Russ.).

Acknowledgments: The article was prepared within the framework of the state task "The Russian legal system in the realities of digital transformation of society and the state: adaptation and prospects for responding to modern challenges and threats (FSMW-2023-0006)". Registration number: 1022040700002-6-5.5.1.

Введение

Буквально за несколько месяцев развитие нейронных сетей вышло на принципиально новый уровень. Анализируя последние достижения искусственного интеллекта (далее – ИИ), мы видим, что ИИ является главным инструментом для автоматизации рутинных, монотонных и сложных задач, которые раньше зависели от ручной работы чело-

века. Технологии искусственного интеллекта позволяют оптимизировать работу сложных производств и уменьшить их затраты. Нейросети могут помочь, например, в создании самого разнообразного контента: от генерации любых изображений и видео по запросам пользователей до подготовки презентации на заданную тему и написании серьёзной научной работы, как, например, случай с защитой диплома студентом РГГУ, созданного при помощи нейросети ChatGPT[1].

Прогнозируя противоправное применение дипфейк-технологий, мы берём в расчёт тот факт, что основной сферой применения генеративного контента является в первую очередь индустрия развлечений. Данный факт, однако, не умаляет криминогенный потенциал дипфейков, а с технологической точки зрения даже является стимулом к развитию его возможностей.

Результаты исследования

Дипфейк представляет собой один из продуктов технологий искусственного интеллекта для создания реалистичных изображений и фонограмм, видеофонограмм, проверка подлинности которых на современном этапе может быть затруднительна.

В настоящее время в российском законодательстве отсутствует нормативное определение понятия «дипфейк». В научных же работах приводятся крайне противоречивые определения.

Например, некоторые авторы используют следующее определение дипфейка: «*Deepfake (Дипфейк, от Deep learning — глубинное обучение и Fake — подделка) является синтезом изображения, основанным на искусственном интеллекте*» [7].

В данном определении понятие «дипфейк» ограничено только генерацией изображений, хотя с технологической точки зрения, дипфейк может принимать самые разные формы, что будет далее описано. И это не единственное определение, которое крайне ограничивает формы дипфейка.

Кроме того, не совсем верно указывается результатом какой деятельности является дипфейк. Дипфейки не ограничиваются лишь синтезом. В определенной степени, в рамках данного понятия можно говорить, например, о синтезе звучащей речи. Однако для всех форм дипфейка следует говорить о генерации цифрового продукта, причём не «основанным на искусственном интеллекте», а созданным при помощи конкретных нейросетевых технологий. Безусловно, нейросетевые технологии входят в понятие искусственный интеллект, однако содержание ИИ ими не ограничивается, поэтому в определении основного понятия считаем необходимым конкретизировать вид технологии на основе ИИ.

Например, в указе Президента РФ [5] приводится следующее определение ИИ, а также описаны технологии, которые он охватывает:

«*а) искусственный интеллект – комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение и поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека. Комплекс технологических решений включает в себя информационно-коммуникационную инфраструктуру, программное обеспечение (в том числе в котором используются методы машинного обучения), процессы и сервисы по обработке данных и поиску решений;*

б) технологии искусственного интеллекта – технологии, основанные на использовании искусственного интеллекта, включая компьютерное зрение, обработку естественного языка, распознавание и синтез речи, интеллектуальную поддержку принятия решений и перспективные методы искусственного интеллекта»;

в) перспективные методы искусственного интеллекта – методы, направленные на создание принципиально новой научно-технической продукции, в том числе в целях разработки универсального (сильного) искусственного интеллекта (автономное решение различных задач, автоматический дизайн физических объектов, автоматическое машинное обучение, алгоритмы решения задач на основе данных с частичной разметкой и (или) незначительных объемов данных, обработка информации на основе новых типов вычислительных систем, интерпретируемая обработка данных и другие методы)».

Изначально дипфейк-технологии создавались для использования в индустрии развлечений: съемок фильмов, рекламных роликов, создании компьютерных игр, однако, как и любые другие технологии, существование дипфейков открывает огромные возможности для достижения противоправных целей.

Для того чтобы обобщить цели использования дипфейков, опишем потенциальные угрозы их применения, среди которых, по нашим оценкам, наибольшего внимания на современном этапе заслуживают следующие:

1. Осуществление шантажа в целях нанесения репутационного вреда, а также вымогательство. С использованием дипфейк-технологий возможно имитировать практически любое событие с реальным субъектом.

В данную группу также следует отнести такое деяние как «порноместь» (с англ. revenge porn). Порнографические видео с реальноживущими субъектами (знаменитостями, политиками и другими людьми) создаются, конечно же, без согласия данных субъектов. Безусловно, распространение подобных видео является грубейшим нарушением на право частной жизни и может нанести серьезный ущерб репутации жертвы [см., например, 2; 8].

Кроме порнографических видео, злоупотребление дипфейк-контентом может создать и любую другую ситуацию, в зависимости от целей злоумышленников. Возможно, сымитировать антиобщественное поведение публичной персоны, например, создать видео с употреблением наркотиков,

[1] Российская газета [Электронный ресурс]. URL: https://rg.ru/2023/02/01/v-rggu-schitaiut-celesoobraznym-ogranichit-dostup-k-nejroseti-chatgpt.html?ysclid=lilhjzfg21119963898 (дата обращения: 01.10.2023).

экстремистскими заявлениями и иными противоправными действиями жертв.

Подобный фейковый контент с постыдными или противоправными действиями будет использован для шантажа и вымогательства, введения аудитории в заблуждение.

2. Мошенничество, в том числе телефонные мошенничества. Для этого преступники уже сейчас клонируют голоса близких родственников потенциальных жертв, а потом от имени данных родственников осуществляют звонки с целью осуществления различных финансовых транзакций с целью хищения денежных средств.

В 2019 году в ходе телефонного разговора генеральный директор крупной энергетической компании перевел почти $250 тыс. мошеннику, голос которого был сгенерирован ИИ и имитировал не только манеру общения руководителя головного офиса, но и легкий немецкий акцент[1].

Например, в мае 2023 года в северном Китае технологии дипфейк использовались злоумышленником для того, чтобы выдать себя за друга жертвы и, таким образом, ввести его в заблуждение во время видеозвонка и обманным путем получить перевод в размере 4,3 миллиона юаней (около 50 миллионов российских рублей)[2].

Кроме того, создаются дипфейки реальных людей с целью имитации их биометрической информации, что позволяет осуществлять в том числе банковские операции по переводу денег. Или заключать различные договоры (купля-продажа собственности, акций, иные сделки, связанные с извлечением какой-либо выгоды). В этом случае технологии дипфейк используются не только для введения в заблуждения отдельных лиц, но и для преодоления пользователем систем контроля и управления доступом.

Распространенным способом мошенничества с использованием дипфейков становится так называемое «криптомошенничество» (с англ. cryptfraud). Злоумышленники создают дипфейки знаменитостей для продвижения розыгрышей криптовалют, от лица известной личности просят перевести криптовалюту и обещают многократную прибыль. Таким образом, вводя пользователей в заблуждение, получают значительные суммы денежных средств.

3. Использование дипфейк-контента в целях манипуляции и пропаганды. Особое место дипфейки занимают в политическом дискурсе. Контент, созданный с помощью дипфейк-технологий, может быть использован в целях оказания влияния на выборы, экономические процессы, в частности на цены товаров или, например, цену акций.

Дипфейки могут создаваться с целью дискредитации политика в рамках предвыборной кампании. Например, злоумышленники могут подготовить и написать для политика заведомо невыгодный для него текст, который будет опубликован с помощью цифрового аватара политика. Естественно, это может нанести существенный вред деловой репутации лица, подрывать общественное доверию к политическим субъектам, а также оказывать влияние на выборы.

Например, в Индии дипфейк использовали для «позитивной» избирательной кампании. Был создан дипфейк, где президент Партии Бхаратия Джаната (БДП) Маноджа Тивари говорит на хинди на диалекте хариани, хотя в оригинале он говорил на английском языке, это было сделано для охвата различных языковых групп избирателей[3].

Таким образом, электорат был введен в заблуждение, хотя сами участники данной избирательной кампании заявляли исключительно о благих целях создания дипфейка.

4. Фальсификация доказательств в судопроизводстве. Дипфейки могут быть использованы для доказывания невиновности преступников или для возбуждения уголовных дел в отношении невиновных.

Не все существующие дипфейк-технологии позволяют генерировать контент высокого качества, однако иногда это предоставляет преимущество злоумышленникам. Как в уголовном производстве, так и в гражданском или производстве по делам об административных правонарушениях доказательствами тех или иных фактов служат, например, видео с камер видеонаблюдения. Качество сигналограмм с камер видеонаблюдения часто низкое, особенно если съемка производилась в темное время суток. С использованием технологий ИИ создать подобное видео становится все проще, никакие специальные знания для этого не требуется. Например, в целях получения страховых выплат собственник может создать фейковое видео, на котором видно, как никогда не существовавшие злоумышленники поджигают склад, наносят повреждение автотранспорту, уносят крупногабаритное имущество, которое в реальности никогда не существовало.

Низкое качество подобных материалов (по причине повсеместного распространения дешевой низкокачественной видеозаписывающей техники) не вызовет подозрения, а проверка обстоятельств, зафиксированных на сигналограмме, судебно-экспертным путем будет существенно затруднена или невозможна.

Таким образом, суд и иные и участники процесса будут введены в заблуждение с использованием современных и широко распространенных технологий.

Важно отметить, что с другой стороны аргументация, связанная с возможностью создания дипфейк-материалов, в ближайшее время станет основной линией защиты по делам, где доказательством выступают цифровые файлы (аналогичная ситуация уже ранее складывалась в отечественном судопроизводстве, когда стороны повсеместно ссылались на изменение данных фотоизображений

[1] Сетевое издание «Коммерсантъ» [Электронный ресурс]. URL: https://www.kommersant.ru/doc/4081979 (дата обращения: 30.10.2023).

[2] Reuters [Электронный ресурс]. URL: https://www.reuters.com/technology/deepfake-scam-china-fans-worries-over-ai-driven-fraud-2023-05-22/ (дата обращения: 30.10.2023).

[3] Портал Хабр [Электронный ресурс]. URL: https://habr.com/ru/news/489006/ (дата обращения: 30.10.2023).

в фотошопе или монтаж компрометирующих их видеоматериалов в видеоредакторах).

5. Дезинформация и фейковые новости для продвижения СМИ и повышения рейтинга цитируемости.

Распространение фейкового контента изначально может и не преследовать противоправных целей, но его распространение с высокой степенью вероятности приведет к негативным последствиям.

Дипфейк-контент может ввести пользователя в заблуждение, даже если изначально создатель таких продуктов не создавал их в противоправных целях.

Например, в мае 2023 года власти Китая сообщили о первом аресте из-за фейковых новостей, генерируемых ChatGPT, когда был задержан мужчина, который с помощью чат-бота ChatGPT создал историю про крушение поезда, в ходе которого погибли люди, и распространил эту новость на китайской платформе для блогов[1]. Мужчине были предъявлены обвинения по статье «провокация конфликтов и беспорядков».

6. Кибербуллинг и преследование с целью провокации агрессии в адрес конкретной личности, чей дипфейк был создан.

Практике уже известны многочисленные случаи актов агрессии, вызванных публикациями в сети интернет. Потенциально возможными представляются ситуации, когда дипфейк-контент может быть использован для провокации агрессии в адрес лица, чья внешность и речь запечатлены в файле, распространяемом для цели кибербуллинга.

7. Нарушение авторских прав.

Пользователи сервисов по созданию дипфейков активно используют их для создания музыкальных треков, написанных от имени популярных исполнителей. Нейросети позволяют клонировать как голос исполнителя, так и сгенерировать текст песни, обучаясь на произведениях того или иного исполнителя. Подобный контент вводит аудиторию в заблуждение, кроме того, правообладатели могут потерять часть дохода от реальных записей таких исполнителей. Например, Лейбл Universal Music потребовал удалить со стримингов музыку, написанную нейросетями[2].

Как мы видим, перечень угроз, связанных с распространением и использованием дипфейков крайне широк. Исходя из этого, мы не находим оснований согласиться с тезисом, что «само по себе создание и распространение дипфейков не обладает какой-либо общественной опасностью» [3]. С учётом приведённого выше перечня угроз следует говорить о двух главных целях распространения и использования дипфейков: **введение в заблуждение или преодоление пользователем систем контроля и управления доступом.**

Что касается классификации дипфейков, то нам представляется целесообразной следующая классификация дипфейк-контента по видам его представления (см. таблицу 1).

Таблица 1. Виды представления дипфейк-контента в цифровой форме

Вид контента	Текст	Графика	Звук
Текст	*статья в СМИ, пост в соцсети и проч.*	*статья в СМИ, пост в соцсети с иллюстрациями (фотографиями фейкового события)*	*синтез звучащей речи / клонирование голоса (обучение нейросети не только на образцах голоса, но и на основе содержательной стороны речи)*
Графика	*статья в СМИ, пост в соцсети с иллюстрациями (фотографиями фейкового события)*	*видео, частично (например, DeepFaceLab) сгенерированные нейросетью, и изображения (например, при помощи Midjourney), полностью сгенерированные нейросетью*	*«цифровой аватар», генерируемый нейросетью типа «Synthesia»*
Звук	*синтез звучащей речи / клонирование голоса (обучение нейросети не только на образцах голоса, но и на основе содержательной стороны речи)*	*«цифровой аватар», генерируемый нейросетью типа «Synthesia»*	*синтезированная речь/клонирование голоса (телефонное мошенничество, озвучивание подготовленного злоумышленниками текста синтезированным или клонированным голосом)*
Комбинация из более чем двух видов	*одновременное использование текста (формализованных параметров устной речи), графики (модель внешности) и звука (клонирование голоса) конкретного человека*[3]		

[1] Российская газета [Электронный ресурс]. URL: https://rg.ru/2023/05/10/chat-bot-dovel-do-tiurmy.html?ysclid=li4mle76h399444598 (дата обращения: 30.10.2023).

[2] Сетевое издание «Коммерсантъ» [Электронный ресурс]. URL: https://www.kommersant.ru/doc/5928221?ysclid=li8vyw9seh543731416 (дата обращения: 30.10.2023).

[3] Например, caryn.ai [Электронный ресурс]. URL: https://caryn.ai (дата обращения: 01.11.2023).

В дополнение к приведённой видовой классификации дипфейк-контента необходима ещё и типовая.

Уже сейчас по типу генерируемого контента можно выделить «неперсонифицированный» – то есть синтезированный с целью создать новый, ранее не существовавший объект и «персонифицированный» контент – то есть результат цифрового воспроизведения конкретно существующего объекта на основе образцов.

Примером неперсонифицированного контента будет являться синтез изображения внешности никогда не существовавшего человека для создания фейкового профиля в социальной сети с целью последующего распространения какой-либо запрещенной законом информации. Подобный пример достаточно простой ситуации может получить последующее развитие, когда профиль в социальной сети будет в последующем дополняться фотоизображениями, сгенерированными на основе ранее синтезированных признаков внешности, сообщениями (постами), имеющими внешнее сходство между собой (по признакам внешности, по особенностям изложения письменной речи).

Исследование такого контента ранее разработанным инструментарием криминалистической диагностики (например, изучение признаков внешности по правилам судебно-портретной экспертизы; диагностика речи автора по правилам автороведческой экспертизы) в целях выявления, раскрытия и расследования преступлений завёл бы правоприменителя в тупик.

Второй тип контента – персонифицированный, может быть использован злоумышленниками для генерации противоправного или компрометирующего контента от имени другого лица. Примером такой ситуации может быть компрометация политического деятеля путем публикации контента (как мы указывали ранее, в виде текстового сообщения, фотографического изображения с комментарием в социальной сети, видеообращения), способного нанести вред его деловой репутации.

Объектный состав такого контента может быть представлен не только изображениями людей, но и предметов и участков местности/помещений, например, когда графический материал изображает компрометируемое лицо в неблагопристойной обстановке.

Важным аспектом предмета правового регулирования с учётом технологических особенностей создания дипфейков должны стать исходные данные, используемые для генерации дипфейк-контента.

Недоступные ранее объёмы образцов голоса, речи и внешности в связи с вынужденным массовым переходом на дистанционную форму осуществления профессиональной деятельности могут в любую минуту стать доступными для злоумышленников, которые, таким образом, получат доступ к образцам, необходимым для синтеза.

Интервью, видеолекции, материалы видеоконференций и другие материалы являются всем необходимым для создания различных форм дипфейков.

Проблема ограничения доступа к таким материалам является особенно острой. Так, например, видеоконференции уже сейчас подвергаются атакам, связанным с осуществлением неправомерного доступа и попытками срыва мероприятий. В научной литературе подобные акты уже получили название zoombombing [10] (зумбомбирование) по названию технологии, получившей широкое распространение во время пандемии коронавируса.

Если раньше целью хакерских взломов становились, например, файлы личных фотографий [9], то новой целью могут стать образцы голоса, речи и внешности, массово доступные в материалах видеоконференций.

Подобные образцы с использованием современных технологий уже сейчас могут быть использованы для совершения преступлений, связанных с противоправным введением в заблуждение при помощи создания дипфейков.

Похожие ситуации возникают по делам об анонимных звонках с сообщениями о готовящихся террористических актах, при требовании выкупа (похищение человека, незаконное лишение свободы и др.) или, например, как мы указывали ранее, в процессе совершения телефонного мошенничества.

Прошлый опыт показывает, что одним из способов сокрытия личности достаточно давно стал синтез звучащей речи. Злоумышленники используют данные технологии для маскировки собственного голоса, когда способ совершения преступления включает вербальный компонент.

В случае генерации персонифицированного дипфейка фонограммы с записью синтезированной речи могут использоваться, например, для имитации коммуникативной ситуации с целью для оговора невиновных. Такого рода материалы уже давно (путем монтажа, а не синтеза) используются злоумышленниками, например, «по делам о получении/даче взятки, об изготовлении и распространении наркотических средств.

Кроме того, маскировка голоса для сокрытия личности используется преступниками при совершении анонимных звонков (заведомо ложное сообщение о террористических актах, преступления против свободы личности, такие как похищение человека, незаконное лишение свободы)» [4, с. 57].

Маскировка голоса и речи используется и для осуществления правосудия. Примером такой деятельности является использование систем синтеза речи для защиты свидетелей при даче ими показаний. Так, например, прибор STC-H773[1] предназначен для использования в системах оснащения залов судебных заседаний с целью защиты тайного свидетеля.

Устройство предназначено для трансформации голоса диктора двумя способами:

– трансформацией голоса при вводе сигнала через микрофон;

– с помощью технологии синтеза речи.

[1] Центр речевых технологий [Электронный ресурс]. URL: https://www.speechpro.ru/product/sistemy-audio-i-videoprotokolirovaniya/stc-h773/specification (дата обращения: 01.11.2023).

Вовлечение данных технологий в судопроизводство также означает, что преступники возьмут подобные технологии на вооружение, но только в противоправных целях.

С учетом сущности дипфейк-технологий, классификации дипфейков по видам и, учитываю перечень потенциальных угроз, предлагаем следующее определение дипфейков: **Дипфейк** – цифровой продукт в виде текста, графики, звука или их сочетания, сгенерированный полностью или частично при помощи нейросетевых технологий, для цели введения в заблуждение или преодоления пользователем систем контроля и управления доступом.

В ситуации отсутствия системы институтов правового регулирования подобных отношений первым этапом, очевидно, потребуется разработка системы принципов, которые необходимо имплементировать в профильное законодательство.

Так, например, проблематика осуществления правосудия с использованием систем искусственного интеллекта нашла отражение в тексте Европейской этической хартии об использовании искусственного интеллекта в судебных системах и окружающих их реалиях, принятая на 31-м пленарном заседании ЕКЭП (Страсбург, 3-4 декабря 2018 года).

В Хартии сформулированы пять принципов:

Принцип уважения основополагающих прав: обеспечить разработку и внедрение инструментов и услуг, основанных на искусственном интеллекте, соответствующих основным правам.

Принцип недискриминации: определенным образом препятствовать развитию или усилению любой дискриминации между отдельными лицами или группами лиц.

Принцип качества и безопасности: при обработке судебных решений и данных, необходимо использовать сертифицированные источники и нематериальные данные с применением моделей, разработанных на междисциплинарной основе, в безопасной технологической среде.

Принцип прозрачности, беспристрастности и достоверности: сделать методы обработки данных доступными и понятными, разрешить проведение внешнего аудита.

Принцип контроля пользователем: избежать предписывающего подхода и позволить пользователю выступать в роли информированного лица, ответственного за свой выбор.

В корпоративной практике также идет процесс выработки принципов, регулирующих деятельность, связанную с применением систем ИИ и их результатами.

Так, компанией Google[1] сформулированы основные цели использования технологий ИИ:

1. Быть социально полезными.

2. Избегать создания или усиления несправедливых предубеждений.

3. Быть построенными и проверенными на безопасность.

4. Быть подотчетным людям.

5. Включать принципы проектирования конфиденциальности.

6. Поддерживать высокие стандарты научного мастерства.

7. Быть доступными для использования в соответствии с этими принципами.

Как мы видим, и здесь делает акцент на безопасности для пользователя и недопустимости каких-либо несправедливых предубеждений (в том числе расовых, религиозных, социальных).

Эффективным инструментом контроля за распространением и использованием дипфейков могут стать включение в тело файлов, генерируемых нейросетями ватермарок (от англ. watermark) и дополнительной служебной информации в метаданных.

О ватермарках, как средстве экспертной профилактики, мы уже ранее упоминали в своих работах:

«*В рамках экспертной профилактики уже сейчас можно предложить:*

регламентацию применения технологии «водяных знаков» (audio watermarking) для маркирования результатов применения программ синтеза речи. Алгоритмы синтеза звучащей речи должны в обязательном порядке автоматически маркировать фонограммы. Подобные технологические особенности возможно урегулировать только на уровне международной отраслевой стандартизации;

международно-правовое регулирование процедур маркирования фонограмм (на этапах разработки, реализации и профессионального применения специализированных программ) для целей противодействия преступлениям, связанным с использованием технологии синтеза звучащей речи» [1].

Watermarks (ватермарки, перевод с англ. водяные знаки) представляют собой определенные цифровые подписи, которые встраиваются в цифровые файлы, например изображения, видеофонограммы и фонограммы. Они могут использоваться для идентификации источника происхождения цифрового продукта, предотвращения несанкционированного копирования или распространения, а также для отслеживания перемещения контента.

Ватермарки действительно могут быть эффективным средством для борьбы с дипфейками. Они должны встраиваться в сгенерированный при помощи технологий ИИ цифровой продукт, вне зависимости от его вида. В разных сферах (в том числе для сферы судопроизводства) подобные технологии уже используются.

Компания Hour One также размещает водяной знак «AV» внизу своих видеофонограмм цифровых аватаров, что означает «Измененные визуальные эффекты»[2].

Компания EDIC-mini, разработчик диктофонов, снабжает файлы с фонограммами, записанными на их

[1] Google [Электронный ресурс]. URL: https://ai.google/responsibility/principles/ (дата обращения 05.11.2023).

[2] GPT-powered deepfakes are a ‘powder keg’ [Электронный ресурс]. URL: https://www.fastcompany.com/90853542 /deepfakes-getting-smarter-thanks-to-gpt (дата обращения: 04.11.2023).

диктофоны, определенными цифровыми маркерами (с дополнительными метаданными), что позволяет защитить запись от несанкционированного использования и осуществить проверку целостности файлов.

Технологически подобные решения уже возможны, поэтому вопрос их правовой регламентации является сугубо техническим.

Существует два основных типа водяных знаков: «хрупкие» и «прочные». Хрупкие цифровые водяные знаки легко разрушаются при манипуляциях с носителем. «Прочные» водяные знаки более устойчивы к манипуляциям, но их сложнее обнаружить.

При регламентации использования цифровых водяных знаков законодателю следует принять во внимание практические сложности, которые могут возникнуть вследствие их использования ватермарок для обнаружения и дипфейков:

– нормы о маркировании дипфейк-контента должны предусматривать применение наиболее защищенных от удаления технологических решений сферы цифровых водяных знаков.

– весьма востребованным представляется вопрос о выявлении цифровых водяных знаков без использования специальных знаний (для тех сфер, которые требуют оперативной проверки контента).

– экономико-правовые аспекты реализации технологии цифрового маркирования и распознавания цифровых водяных знаков.

– сложности антимонопольного регулирования деятельности по генерации дипфейк-контента с учетом массового характера внедрения таких технологических решений.

В конце 2021 г. в России направление, связанное с выявлением дипфейков включили в дорожную карту «Новые коммуникационные интернет-технологии» [6]. Этот документ был разработан «Ростелекомом» и утвержден Правительственной комиссией по цифровому развитию.

Примеры регулирования можно отследить и в индустрии.

Коалиция за происхождение и подлинность контента (The Coalition for Content Provenance and Authenticity) развивает Инициативу Adobe по аутентичности контента (CAI) (The Content Authenticity Initiative), предлагая пользователям инструменты, которые позволяют интегрировать в цифровые продукты так называемые «безопасные сигналы происхождения». Эти инструменты позволяют легко указать, когда ИИ использовался для создания или изменения контента. Информация о конкретных использованных моделях ИИ и другая информация может быть передана пользователям, что помогает предотвратить дезинформацию и повысить прозрачность процесса и результатов использования ИИ[1], снижая риски его противоправного применения.

Интересным представляется опыт Китая. Китайское правительство в конце 2022 года приняло Положение об управлении глубоким синтезом информационных услуг Интернета, в соответствии с которым разработчики генеративного контента (дипфейков) должны добавлять определенные логотипы (watermarks, цифровые водяные знаки) в файлы-результаты генерации. Это касается чатботов, синтезированной речи, генерации изображений и видео лиц и т.д.[2].

Европарламент уже начал обсуждение законопроекта, который направлен на регулирование применения ИИ. Законопроект уже опубликован на сайте Европарламента[3]. Его принятие планируется на конец 2023 или начало 2024 года.

Некоторые механизмы правового регулирования могут быть взяты, например, из сферы применения блокчейн-технологии. Так, например, специализированные ресурсы[4] прогнозируют технологические решения, которые, по нашему мнению, могут быть имплементированы в отечественное законодательство.

В качестве механизма проверки того, что видео является оригиналом, по мнению экспертов, может быть использована технология кодирования видео в неизменяемые блокчейны (в индустрии подобные технологии реализуют такие компании, как Factom, Ambervideo[5] и Axiom).

«*Основная идея многих таких проектов заключается в том, что данные, содержащиеся в видеофайле или сгенерированные определенной камерой, могут использоваться для создания уникальной подписи, которая изменится, если видео будет изменено. В конце концов, видео, загруженные в социальные сети, могут иметь генерировать код аутентификации, который исходный загрузчик может зарегистрировать в блокчейне, чтобы доказать, что они являются первоначальными владельцами видео.*

Эти решения, конечно, имеют свой собственный набор проблем, таких как кодирование видео, изменяющее данные в файле и изменение подписи без фактического изменения видеоконтента, или законное редактирование видео, искажающее подпись. Однако, в ситуациях с высокими ставками, таких как коммерческие транзакции, когда изображения используются для проверки доставки или получения поддержки инвестора, наличие такого уровня аутентификации может помочь предотвратить мошенничество, связанное с дипфейком»[6].

[1] Contentauthenticity.org [Электронный ресурс]. URL: https://contentauthenticity.org (дата обращения: 04.11.2023).

[2] Государственная канцелярия интернет-информации КНР (англ. Cyberspace Administration of China) [Электронный ресурс]. URL: http://www.cac.gov.cn/2022-12/11/c_1672221949318230.htm (дата обращения: 04.11.2023).

[3] [Электронный ресурс]. URL: https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/CJ40/DV/2023/05-11/ConsolidatedCA_IMCOLIBE_AI_ACT_EN.pdf (дата обращения: 04.11.2023).

[4] Deepfakechallenge [Электронный ресурс]. URL: https://deepfakechallenge.com/ (дата обращения: 01.11.2023).

[5] Ambervideo [Электронный ресурс]. URL: https://ambervideo.co/ (дата обращения: 01.11.2023).

[6] Deepfakechallenge [Электронный ресурс]. URL: https://deepfakechallenge.com/2021/03/15/7953/ (дата обращения: 01.11.2023).

Подобно рода технологии на начальном этапе могут быть имплементированы в процессуальное законодательство, например для формальной оценки цифровых файлов судом. Во всяком случае подобные решения представляются реальными, применимыми и востребованными правоприменительной практикой.

Заключение

Не стоит говорить о полном запрете на создание генеративного контента или пытаться добавить в каждую статью норму об использовании дипфейк-технологий как отягчающего обстоятельства. Необходимо прорабатывать механизмы предупреждения использования и распространения дипфейков, закрепив нормативно определение «дипфейка», вводя нормы об обязательной маркировке дипфейков и использовании водяных знаков при их создании, разработав, таким образом, общеправовой запрет на распространение и использование дипфейков без специальной маркировки.

Таким образом, нам необходимо должным образом осмыслить нормативное регулирование данного вопроса с учётом правовых пробелов, которые сопровождают современный уровень искусственного интеллекта, и потенциальных угроз, которые за последнее время становятся все более реальными.

Список источников

1. *Бодров Н. Ф., Лебедева А. К.* Перспективы судебно-экспертного исследования синтезированной звучащей речи // Законы России: опыт, анализ, практика. 2021. № 3. С. 9-13.

2. *Грачева Ю.В.* Цифровые технологии и безопасность личности // Криминалистъ. 2019. № 1. С. 24-27.

3. *Грешнова Н. А.*, *Ситник В. Н.* Обеспечение общественного интереса в условиях цифровизации: проблемы уголовного законодательства в России (на примере технологии дипфейк (Дипфейк)) // Вестник Саратовской государственной юридической академии. 2022. № 5(148). С. 182-189.

4. *Лебедева А.К.* Технологии голосового синтеза и судебная фоноскопическая экспертиза // Вестник криминалистики. 2020. № 3 (75). С. 57.

5. Национальная стратегия развития искусственного интеллекта в России до 2030 г., утвержденная Указом Президента Российской Федерации от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации». Доступ из справ.-правовой системы «КонсультантПлюс».

6. О государственной поддержке программ деятельности лидирующих исследовательских центров, реализуемых российскими организациями в целях обеспечения разработки и реализации дорожных карт развития перспективных «сквозных» цифровых технологий: постановление Правительства РФ от 03.05.2019 № 551 (ред. от 19.12.2019). Доступ из справ.-правовой системы «КонсультантПлюс».

7. *Ситник В. Н.* Перспективы установления уголовной ответственности за преступления, совершенные с использованием технологии дипфейк // Уральский журнал правовых исследований. 2022. № 3(20). С. 76-83.

8. *Юрченко И.А.* Сталкер как субъект уголовной ответственности // Вестник Университета имени О. Е. Кутафина. 2018. № 12 (52). С. 53-61.

9. *Marwick A.E.* Scandal or sex crime? Gendered privacy and the celebrity nude photo leaks // Ethics and Information Technology. 2017. Т. 19, № 3. С. 177-191.

10. *Walsh C.G., Unertl K.M., Ebert J.S.* Rapid Supportive Response to a Traumatic “Zoombombing” During the COVID-19 Pandemic // Academic Medicine. 2021. Т. 96, № 1. С.6-7.

References

1. Bodrov N. F., Lebedeva A. K. Perspektivy sudebno-ekspertnogo issledovaniya sintezirovannoj zvuchashchej rechi [Prospects for forensic investigation of synthesized sounding speech], *Zakony Rossii: opyt, analiz, praktika [Laws of Russia: experience, analysis, practice]*, 2021, no. 3, pp. 9-13. (In Russ.).

2. Gracheva YU.V. Cifrovye tekhnologii i bezopasnost' lichnosti [Digital technologies and personal security], *Kriminalist [Criminnalist]*, 2019, no. 1, pp. 24-27. (In Russ.).

3. Greshnova N. A., Sitnik V. N. Obespechenie obshchestvennogo interesa v usloviyah cifrovizacii: problemy ugolovnogo zakonodatel'stva v Rossii (na primere tekhnologii dipfejk (Dipfejk)) [Ensuring public interest in the conditions of digitalization: problems of criminal legislation in Russia (on the example of deepfake technology (deepfake)], *Vestnik Saratovskoj gosudarstvennoj yuridicheskoj akademii [Bulletin of the Saratov State Law Academy]*, 2022, no. 5(148), pp. 182-189. (In Russ.).

4. Lebedeva A.K. Tekhnologii golosovogo sinteza i sudebnaya fonoskopicheskaya ekspertiza [Voice synthesis technologies and forensic phonoscopic examination], *Vestnik kriminalistiki [Vestnik of criminnalistics]*, 2020, no. 3 (75). (In Russ.).

5. Nacional'naya strategiya razvitiya iskusstvennogo intellekta v Rossii do 2030 g., utverzhdennaya Ukazom Prezidenta Rossijskoj Federacii ot 10.10.2019 № 490 "O razvitii iskusstvennogo intellekta v Rossijskoj Federacii" [National Strategy for the Development of Artificial Intelligence in Russia until 2030, approved by Decree of the President of the Russian Federation dated 10.10.2019 No. 490 "On the development of artificial intelligence in the Russian Federation"]. Access from the help-legal system "ConsultantPlus". (In Russ.).

6. O gosudarstvennoj podderzhke programm deyatel'nosti lidiruyushchih issledovatel'skih centrov, realizuemyh rossijskimi organizaciyami v celyah obespecheniya razrabotki i realizacii dorozhnyh kart razvitiya perspektivnyh "skvoznyh" cifrovyh tekhnologij: Postanovlenie Pravitel'stva RF ot 03.05.2019 N 551 (red. ot 19.12.2019) [On state support for programs of activity of leading research centers implemented by Russian organizations to ensure the development and implementation of roadmaps for the development of promising "end-to-end" digital technologies: Resolution of the Government of the Russian Federation dated 05.03.2019 No. 551 (as amended on 12.19.2019)]. Access from the help-legal system "ConsultantPlus". (In Russ.).

7. Sitnik V. N. Perspektivy ustanovleniya ugolovnoj otvetstvennosti za prestupleniya, sovershennye s ispol'zovaniem tekhnologii dipfejk [Prospects of establishing criminal liability for crimes committed with the use of deepfake technology], *Ural'skij zhurnal pravovyh issledovanij [Ural Journal of Legal Studies]*, 2022, no. 3(20), pp. 76-83. (In Russ.).

8. YUrchenko I.A. Stalker kak sub"ekt ugolovnoj otvetstvennosti [Stalker as a subject of criminal responsibility], *Vestnik Universiteta imeni O. E. Kutafina [Bulletin of Kutafin Moscow State Law University]*, 2018, no. 12 (52), pp. 53-61. (In Russ.).

9. Marwick A.E. Scandal or sex crime? Gendered privacy and the celebrity nude photo leaks. *Ethics and Information Technology*, 2017, vol. 19, no. 3, pp. 177-191.

10. Walsh C.G., Unertl K.M., Ebert J.S. Rapid Supportive Response to a Traumatic “Zoombombing” During the COVID-19 Pandemic. *Academic Medicine*, 2021, vol. 96, no. 1, pp. 6-7.

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

Authors' contribution: All authors have made an equivalent contribution to the preparation of the publication. The authors declare that there is no conflict of interest.

ИНФОРМАЦИЯ ОБ АВТОРАХ

Бодров Николай Филиппович, кандидат юридических наук, доцент кафедры судебных экспертиз Университета имени О.Е. Кутафина (МГЮА), г. Москва, Российская Федерация. *E-mail: zhuravlenok_os@mail.ru*

Лебедева Антонина Константиновна, кандидат юридических наук, доцент кафедры судебных экспертиз Университета имени О.Е. Кутафина (МГЮА), г. Москва, Российская Федерация. *E-mail: tonya109@yandex.ru*

INFORMATION ABOUT THE AUTHORS

Nikolay F. Bodrov, Phd (Law), Lecturer of the Department of Forensic Expertise’s department. Kutafin Moscow State Law University (MSAL). Moscow, Russian Federation. *E-mail: zhuravlenok_os@mail.ru*

Antonina K. Lebedeva, Phd (Law), Lecturer of the Department of Forensic Expertise’s department. Kutafin Moscow State Law University (MSAL), Moscow, Russian Federation. *E-mail: tonya109@yandex.ru*

Поступила в редакцию 05.11.2023 г.; одобрена после рецензирования 04.12.2023 г.; принята к публикации 05.12.2023 г.

Received 05.11.2023; approved after reviewing 04.12.2023; accepted for publication 05.12.2023