Научная статья
УДК 343.3/.7:004
doi 10.46741/2686-9764.2024.68.4.001

Перспективы правового регулирования и алгоритм маркировки генеративного контента

НИКОЛАЙ ФИЛИППОВИЧ БОДРОВ
Университет имени О. Е. Кутафина (МГЮА), Москва, Россия, bodrovnf@gmail.com, https://orcid.org/0000-0002-9005-3821

АНТОНИНА КОНСТАНТИНОВНА ЛЕБЕДЕВА
Университет имени О. Е. Кутафина (МГЮА), Москва, Россия, tonya109@yandex.ru, https://orcid.org/0009-0004-9344-2103

Реферат

Введение: статья посвящена анализу механизмов правового регулирования генеративного контента в Российской Федерации, алгоритма его маркирования. Анализ действующего законодательства показывает, что возрастание роли технологий искусственного интеллекта в современном обществе диктует необходимость совершенствования существующих, а также разработки новых норм для правового регулирования оборота генеративного контента. Российский законодатель предпринимает первые попытки разработки законодательных мер контроля за оборотом генеративного контента, однако в настоящее время отсутствует даже унифицированное определение дипфейка как основополагающего понятия данной сферы. *Цель*: на основе существующей классификации видов генеративного контента описать существующие виды его маркировки, представить определение дипфейка. *Методы*: сравнительно-правовой, эмпирические методы описания, интерпретации, теоретические методы формальной и диалектической логики, юридико-догматический и метод толкования правовых норм. *Результаты*: анализ видов генеративного контента, целей его распространения демонстрирует необходимость создания эффективных правовых и технологических механизмов регулирования. Разработаны теоретические положения о структуре субъектов создания и распространения генеративного контента, предложены рекомендации по установлению их правовых обязанностей и ответственности. *Выводы*: механизмы правового регулирования генеративного контента на данном этапе сводятся к следующему: закрепление обязанности технологических компаний и пользователей, создающих генеративный контент, использовать водяные знаки для маркирования контента в целях информирования иных лиц о генеративной природе контента; установление ответственности за отказ от маркировки или удаление какого-либо вида маркировки, неправомерное использование биометрических персональных данных для создания дипфейка; учет общественной опасности деяний, совершаемых с использованием дипфейка в качестве средства.

Ключевые слова: дипфейк; генеративный контент; правовое регулирование; маркировка контента; ответственность.

5.1.1. Теоретико-исторические правовые науки.

5.1.2. Публично-правовые (государственно-правовые) науки.

5.1.4. Уголовно-правовые науки.

Благодарность: публикация подготовлена в рамках выполнения работ по государственному заданию на тему: «Российская правовая система в реалиях цифровой трансформации общества и государства: адаптация и перспективы реагирования на современные вызовы и угрозы (FSMW-2023-0006)». Регистрационный номер ЕГИСУНИОКТР: 124012000079-6

Для цитирования: Бодров Н. Ф., Лебедева А. К. Перспективы правового регулирования и алгоритм маркировки генеративного контента // Пенитенциарная наука. 2024. Т. 18, № 4 (68). С. 348–357. doi 10.46741/2686-9764.2024.68.4.001.

Original article

Prospects of Legal Regulation and Algorithm for Marking Generative Content

NIKOLAI F. BODROV
Kutafin Moscow State Law University (MSAL), Moscow, Russia, bodrovnf@gmail.com, https://orcid.org/0000-0002-9005-3821

ANTONINA K. LEBEDEVA,
Kutafin Moscow State Law University (MSAL), Moscow, Russia, tonya109@yandex.ru, https://orcid.org/0009-0004-9344-2103

Abstract
Introduction: the article analyzes mechanisms for legal regulation of generative content in the Russian Federation and considers an algorithm of marking generative content. The analysis of current legislation shows that the increasing role of artificial intelligence technologies in modern society necessitates the improvement of existing norms, as well as the development of new ones for the legal regulation of generative content turnover. The Russian legislator is trying to develop legislative measures to control the turnover of generative content, but currently there is not even a unified definition of a deepfake as a fundamental concept of this sphere. *Purpose*: on the basis of the existing classification of generative content types to describe the existing types of generative content labelling and to present the definition of a deepfake. *Methods*: comparative-legal, empirical methods of description, interpretation; theoretical methods of formal and dialectical logic, legal-dogmatic and method of interpretation of legal norms, are applied. *Results*: the analysis of existing types of generative content, the purposes of its distribution demonstrates the need to create effective legal and technological mechanisms for regulating generative content. Theoretical provisions on the structure of subjects of creation and distribution of generative content are developed and recommendations on establishing their legal duties and responsibilities are offered. *Conclusions*: the mechanisms of legal regulation of generative content at this stage are reduced to the following: establishing the obligation of technology companies and users who create generative content to use watermarks to mark the content in order to inform other persons about the generative nature of the content; establishing liability for the refusal of marking or removal of any type of marking; establishing liability for the misuse of biometric personal data for the creation of a deepfake.

Key words: deepfake, generative content, legal regulation, content marking, liability.

5.1.1. Theoretical and historical legal sciences.

5.1.2. Public law (state law) sciences.

5.1.4. Criminal law sciences.

Acknowledgements: This publication has been prepared as part of the work under the state assignment on the topic “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)”. The EGISUNIOKTR registration number: 124012000079-6.

For citation: Bodrov N.S., Lebedeva A.K. Prospects of legal regulation and algorithm for marking generative content. *Penitentiary Science*, 2024, vol. 18, no. 4 (68), pp. 348–357. doi 10.46741/2686-9764.2024.68.4.001.

Введение

Технологии искусственного интеллекта (ИИ), позволяющие создавать различные виды контента, – мощный инструмент для творческого самовыражения, востребованный в самых разных сферах человеческой деятельности – от искусства до, например,

маркетинга. Однако бесконтрольное распространение различного генеративного контента, созданного при помощи современных нейросетевых алгоритмов, способных синтезировать текст, графику и звуки, уже сейчас приводит к серьезным последствиям, заслуживающим должного внимания законодателя и правоприменителя.

Важнейшей проблемой стал практически бесконтрольный доступ к биометрическим персональным данным лица на основе тех медиаматериалов, которые в последние годы распространялись в открытом доступе в социальных сетях, системах обмена мгновенными сообщениями, сервисах облачного хранения, на файлообменниках, в системах хранения данных программ видео-конференц-связи и видеохостингах. Указанные материалы уже сейчас являются основой для создания цифровых продуктов в виде текста, графики, звука или их сочетания, сгенерированных полностью или частично при помощи нейросетевых технологий для цели введения в заблуждение или преодоления пользователем систем контроля и управления доступом [1]. Подобного рода продукты являются по своей сути содержательным наполнением термина «дипфейк».

Усугубляет рассматриваемую проблематику тот факт, что технологии нейросетевого синтеза дипфейк-контента уже достаточно давно перешли в разряд сервисов по запросу, когда, используя специально настроенные приложения или сайты, пользователь с любым уровнем технической подготовки может пошагово следовать инструкциям и создать дипфейк. Коммерциализация технологий искусственного интеллекта, бесспорно, является серьезной угрозой для информационной безопасности общества и государства.

Так, по данным полиции Гонконга, финансовый служащий выплатил 25 миллионов долларов после видеозвонка с финансовым директором, внешность и голос которого были синтезированы злоумышленником с использованием нейросетевых технологий [2].

Стоит признать и риски ошибок второго рода, когда системы биометрической идентификации будут ошибочно отказывать пользователям, полагаясь на ложные срабатывания детекторов дипфейков. Например, в системе Admitad пользователю было отказано в обслуживании в результате видеоверификации, поскольку система ошибочно классифицировала его внешность как дипфейк [3].

Даже предварительные результаты анализа статистики правонарушений, связанных с распространением дипфейков, показывают весьма негативную динамику. Так, например, ссылаясь на отчет платформы проверки личности Sumsub, исследователи сообщают, что различные инциденты с дипфейками в финансовом секторе выросли на 700 % в 2023 г. по сравнению с предыдущим годом [4].

Если год назад законодатель считал нецелесообразным криминализацию использования технологий искусственного интеллекта в преступных целях [5], то сейчас для российской правовой системы очевидна потребность в создании механизмов правового регулирования генеративного контента [6], в том числе и в качестве квалифицирующего признака [7] преступления. Законопроект предусматривает внесение в некоторые статьи УК РФ дополнительного квалифицированного состава – совершение преступления «с использованием изображения или голоса (в том числе фальсифицированных или искусственно созданных) потерпевшего или иного лица, а равно с использованием биометрических персональных данных потерпевшего или иного лица» [7].

Изменения планируется внести в части таких статей УК РФ, как «Клевета», «Кража», «Мошенничество», «Вымогательство», «Причинение имущественного ущерба путем обмана или злоупотребления доверием» (ч. 2.1 ст. 128.1, п. «д» ч. 3 ст. 158, ч. 2.1 ст. 159, п. «д» ч. 2 ст. 163, п. «в» ч. 2 ст. 165).

Законодатель не дает четкого определения фальсифицированных или искусственно созданных голосов. Однако, исходя из содержания пояснительной записки и отзывов на законопроект, можно предположить, что под ними подразумеваются дипфейки.

Однако перечень статей представляется нам крайне ограниченным [8], список реальных и потенциальных угроз, связанных с использованием и распространением дипфейков, значительно шире.

Так, по нашему мнению, в самое ближайшее время распространение систем генеративного синтеза контента приведет к существенной трансформации преступлений (в первую очередь, способов их совершения):

– против личности (ст. 110, 128.1, 146 УК РФ);

– в сфере экономики (ст. 159, 159.3, 159.6, 185.3 УК РФ),

– против общественной безопасности и общественного порядка (ст. 205.2, 207, 207.1, 207.3, 242, 242.1, 272–274 УК РФ);

– против государственной власти (ст. 280, 280.1, 280.3, 280.4, 282, 282.4, 284.2, 303 УК РФ),

– против мира и безопасности человечества (ст. 354 УК РФ) [9].

Данный прогноз строится на том, что в составах указанных выше преступлений дипфейк наиболее вероятно может быть использован в качестве орудия совершения преступления.

Подход законодателя в части выделения крайне ограниченного перечня статей в указанном выше законопроекте нам представляется не совсем верным. Появление новых квалифицированных составов на данном этапе не решит проблему использования и распространения дипфейков, а первоочередной задачей является создание новых и усовершенствование существующих механизмов регулирования оборота генеративного контента, созданного при помощи искусственного интеллекта.

В отзыве на законопроект сказано, что «в отраслевом законодательстве не урегулированы вопросы использования технологий подмены личности (дипфейк). Таким образом, введение предлагаемого регулирования в уголовное законодательство не представляется возможным из-за отсутствия корреспондирующих норм материального законодательства, что влечет существенные риски формирования некорректной правоприменительной практики» [10].

Таким образом, предлагая защиту граждан от дипфейков, разработчики законопроекта не уточняют, что конкретно подразумевается под термином «дипфейк». Корректность использования термина «искусственная аудиозапись» с позиций судебной экспертологии и судебно-следственной практики вызывает сомнения по причине того, что отражает существенно более широкий объем понятия.

По нашему мнению, термин «дипфейк» необходимо определить следующим образом: дипфейк – это 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. Необходимо закрепить понятие дипфейка в Федеральном законе от 27.07.2006 № 149-ФЗ «Об информации, информационных технологиях и о защите информации».

Одной только имплементации концепции дипфейка в российское законодательство явно недостаточно, поскольку за пределами внимания законодателя остаются вопросы правовой регламентации создания, использования и распространения генеративного контента, а для правоприменения важно четко отграничивать дипфейк от других видов генеративного контента.

Более того, осуществляя научные исследования в сфере терминологического обеспечения описанной проблематики, мы пришли к выводу, что в научной и нормативно-технической литературе в настоящий момент отсутствуют не только определение, но и подходы к описанию такого явления, как «селффэйк» (с англ. self – «себя» или «сам»).

Обобщая результаты оценки рисков противоправного распространения генеративного контента, полагаем, что под селффейком следует понимать дипфейк, сгенерированный пользователем на основе его собственных биометрических данных с целью совершения противоправных действий, направленных на уклонение от ответственности или введение в заблуждение других лиц относительно событий, представленных как происшедших с самим пользователем. Примером таких событий могут быть ситуации страхового мошенничества [11], когда пользователи при помощи алгоритмов нейросетевого синтеза генерируют фотографии с повреждениями имущества для получения страховых выплат. В такой ситуации сотрудники страховых организаций могут вступать в сговор со злоумышленниками и предоставлять вместо фотографий предстрахового осмотра изображения, сгенерированные нейросетью.

Правовые механизмы противодействия распространению и использованию дипфейк-контента должны предоставлять возможность как обычным пользователям, так и лицам, обладающим специальными знаниями, осуществить детекцию такого контента. Если в качестве механизмов обнаружения дипфейков можно предусмотреть меры судебно-экспертного противодействия [12], то мерами предотвращения использования и распространения дипфейков является разработка правовых норм, направленных на урегулирование технологий и процессов введения обязательной маркировки генеративного контента.

Дипфейк является лишь разновидностью генеративного контента. Распространяя такой контент, создатели не скрывают источник его происхождения, даже если результат генерации достигает высокой степени реалистичности. В ходе правомерного использования генеративного контента необходимо обеспечить доступность информации о способе его создания. Распространение же дипфейков связано с сокрытием способа имитации аутентичного контента. Для информирования пользователей о факте генерации контента при помощи технологий искусственного интеллекта необходимо разработать нормы, предусматривающие обязательность эффективного маркирования любого генеративного контента.

В настоящее время уже невозможно предусмотреть хоть сколько-нибудь эффективные меры запрета на создание генеративного контента. В Национальной стратегии развития искусственного интеллекта, утвержденной Указом Президента Российской Федерации от 15.02.2024 № 124, подчеркивается необходимость закрепления благоприятных нормативно-правовых условий для разработки и внедрения технологий искусственного интеллекта, полный же запрет на технологии генеративного искусственного интеллекта таких условий, конечно же, не реализует. Потенциальные и реальные угрозы исходят не от факта создания технологий для генерации различного контента, а при недобросовестном использовании результатов такой генерации без указания способа получения цифрового продукта. Можно с уверенностью утверждать, что существует риск использования и распространения генеративного контента, который легко может быть ошибочно принят за аутентичный.

Способы маркирования генеративного контента

Для упорядочивания оборота генеративного контента без ущерба для информационной безопасности и предотвращения дезинформации необходимо предусмотреть действенные способы маркирования и ограничить его использование без соответствующей маркировки.

В подобной ситуации механизмы правового регулирования фактически сводятся к созданию норм о маркировании генеративного контента с целью поддержания информационной осведомленности пользователей о таком контенте и его происхождении. В разных сферах (в том числе для сферы судопроизводства) пригодные для практического применения технологии уже используются. Законодательство, таким образом, должно ограничивать оборот генеративного контента и устанавливать ответственность за преодоление механизмов информирования о генеративной природе контента.

Вместе с механизмами правового регулирования необходимо также разработать и технологические способы маркировки. В первую очередь возникает потребность в единых международных стандартах в области маркировки генеративного контента. Процесс и технологии распространения дипфейков не имеют географических границ. Генеративный контент создается, используется и распространяется во

всем мире. Следовательно, необходима разработка унифицированных стандартов маркировки генеративного контента, чтобы дать возможность пользователям легко их детектировать, независимо от языка, используемой технологической платформы или государства, на территории которого проживает потребитель контента.

Среди используемых в настоящее время способов достаточно широкое распространение получила графическая маркировка. В качестве маркеров в таких системах выступают краткий текст к какому-либо цифровому продукту (например, «создано при помощи ИИ») по аналогии с маркировкой иноагентов или экстремистских организаций (федеральные законы от 14.07.2022 № 255-ФЗ «О контроле за деятельностью лиц, находящихся под иностранным влиянием», от 25.07.2002 № 114-ФЗ «О противодействии экстремистской деятельности»), или же графические объекты, используемые по аналогии с маркировкой информационной продукции в соответствии с Федеральным законом от 29.12.2010 № 436-ФЗ «О защите детей от информации, причиняющей вред их здоровью и развитию». Такая маркировка носит уведомительный характер, но в силу простой формы представления не может быть использована в специализированных сферах, например, в сфере судопроизводства.

Технологически более сложным вариантом является маркировка с помощью водяных знаков. Ватермарки (с англ. watermarks – водяные знаки) представляют собой 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 [8].

Так, Группа компаний «Телесистемы», существующая на российском рынке более 20 лет, разрабатывает и производит диктофоны EDIC-mini [13]. Файлы с фонограммами, записанными на диктофоны данной марки, снабжаются определенными цифровыми маркерами подлинности записи (с дополнительными метаданными), что позволяет защитить запись от несанкционированного использования и осуществить проверку целостности файлов. Данная функция является крайне востребованной в судебно-следственной и экспертной практике.

Компания «Hour One», создающая так называемые цифровые аватары, использует водяной знак «AV» (Altered Visuals, «измененные визуальные эффекты») [14] для маркирования своих видео. Как указывают разработчики, делается это из уважения права конечного пользователя знать, что видео было сгенерировано с помощью технологий искусственного интеллекта.

Для предотвращения распространения дипфейков и дезинформации крупные корпорации уже имеют опыт подтверждения аутентичности цифровых продуктов. Коалиция за происхождение и подлинность контента (The Coalition for Content Provenance and Authenticity) предлагает определенную маркировку для подтверждения аутентичности контента: «Content Credentials “iconoftransparency”» (стандарт C2PA) [18].

Разработанный коалицией технологический стандарт C2PA позволяет креаторам встраивать метаданные в цифровые продукты для проверки их происхождения и связанной с ними информации. Стандарт предназначен не только для генеративных изображений, его использование планируется и производителями фотокамер, медиакомпаниями, создающими визуальный контент для сертификации источника и происхождения медиаконтента. В свою деятельность интегрировали данный стандарт такие компании, как OpenAI, Adobe, Microsoft, Publicis Groupe, Leica, Nikon.

Подобная маркировка будет содержать как невидимые для глаза пользователя метаданные, хранящиеся в файле, так и видимый водяной знак CR, который будет появляться в левом верхнем углу каждого изображения.

Изображения, сгенерированные при помощи таких нейросетей, как ChatGPT и DALL·E с февраля 2024 г. включают метаданные C2PA. Для проверки истории изображения пользователь может использовать их сервис «Content Credentials Verify» [16]. Однако используемый подход маркировки распространяется только на графические формы генеративного контента. Кроме того, пользователь может удалить маркировку, сделав обычный скриншот сгенерированного изображения, тем самым создается новый файл с чистыми метаданными. Водяной знак также может быть удален при кадрировании изображения.

Более того, компания Microsoft объявила о создании сервиса «Azure OpenAI Service», который добавляет невидимые водяные знаки ко всем изображениям, генерируемым с помощью DALL·E [17]. Подобный сервис существует у компании и при генерации синтезированных голосов при помощи их разработки – Azure AI Speech [18], встраиваемые водяные знаки позволяют определить, была ли синтезирована речь с помощью Azure AI Speech, а также какой голос был использован при генерации.

Компания Google представила бета-версию инструмента для встраивания цифровых водяных знаков непосредственно в изображения, аудио, текст или видео, созданные при помощи искусственного интеллекта, – SynthID [19]. Данный инструмент позволяет пользователям встраивать цифровой водяной знак непосредственно в цифровые продукты, созданные искусственным интеллектом. Впоследствии на основе этих водяных знаков возможно осуществлять верификацию контента, но только если он был сгенерирован с помощью моделей искусственного интеллекта от Google.

Существуют сервисы, предлагающие встраивать водяные знаки для аудиофайлов, чтобы предотвратить использование данных пользователя моделями искусственного интеллекта, отслеживая их обратно к источнику. Свои технологии для создания водяных знаков предлагают и разработчики систем для синтеза и клонирования звучащей речи. Например, компания Resemble AI разработала сложный глубокий нейросетевой водяной знак «PerTh» [20], который по-

зволяет встраивать незаметные данные в генерируемую звучащую речь, создавая невидимый водяной знак. По мнению разработчиков, данный знак трудно удалить, однако возможность проверки имеется только для фонограмм, сгенерированных с помощью Resemble AI.

Нейросеть MyVocalAI [21] предлагает различные возможности для синтеза и клонирования звучащей речи и обеспечивает пользователей возможностью создания водяных знаков для маркирования генеративного контента. Однако пользователи с платной подпиской (от тысячи рублей в месяц) получают техническую возможность удалять маркировку, создавать фонограммы с клонированными голосами, не только непосредственно записывая свой голос, но и загрузив любые фонограммы любого диктора.

В связи с тем что значительный объем генеративного контента распространяется в социальных сетях, эффективным инструментом может стать создание служб на различных социальных платформах, которые автоматически бы распознавали генеративный контент и уведомляли бы пользователей об этом. Несмотря на технологическую трудоемкость данного процесса, некоторые корпорации имеют достаточные ресурсы для создания таких сервисов.

Например, видеохостинговый сервис Youtube (иностранное лицо, владеющее информационным ресурсом, является нарушителем законодательства Российской Федерации) также заявил о новой обязанности пользователей раскрывать информацию о том, является ли загружаемое видео измененным или синтезированным [22]. Кроме того, компания заявила, что их последнее обновление будет также автоматически выявлять подобный контент и ставить данные лейблы в описание к видео. Однако пользователи уже заявляют, что механизм детекции сгенерированного контента пока не отлажен и работает крайне нестабильно, имеют место случаи, когда аутентичный контент был распознан как сгенерированный. Компания также планирует в ближайшие несколько месяцев наладить алгоритм удаления генеративного контента, если пользователи увидят, что для его создания были использованы пользовательские лицо или голос. Однако нельзя исключить случаи злоупотребления данным правом, когда, например, публичные персоны будут подавать заявку на удаление дискредитирующих их видео, сообщая, что данный контент является сгенерированным, в то время как он является аутентичным.

Компания Meta (деятельность организации признана экстремистской и запрещена на территории Российской Федерации) заявила, что планирует маркировать изображения, видео, аудио в своих социальных сетях текстом «Made with AI» (сделано с помощью ИИ), если сами разработанные алгоритмы выявят признаки сгенерированного с помощью нейросетей контента либо если пользователь сам уведомит об этом. Если фотореалистичное изображение сгенерировано с помощью их нейросети Meta AI (принадлежит Meta Platforms Inc – организации, деятельность которой признана экстремистской и запрещена на территории Российской Федерации), оно уже маркируется текстом «Imaginedwith AI» [23].

В июне 2024 г. данная компания представила технологию для встраивания в файлы с фонограммами клонированной речи особых водяных знаков «AudioSeal» [24]. Вместе с технологией для создания водяных знаков представили и средство для детекции данных водяных знаков, что существенно упрощает верификацию клонированного голоса. По заявлению разработчиков, применяемые водяные знаки устойчивы даже в случае различных вариантов редактирования звукозаписи. Водяной знак представляет собой определенный сигнал в сгенерированной фонограмме, который не различим для человеческого уха. Сведения об устойчивости маркировки в ситуации квалифицированного монтажа фонограммы в открытой печати отсутствуют.

AudioSeal представляет собой метод локализованного водяного знака речи, он совместно обучает две сети: генератор, который предсказывает аддитивную форму волны водяного знака на входном аудиосигнале, и детектор, который выдает вероятность наличия в аудиосигнале водяного знака.

Сервис для создания различных видов контента Descript, который также дает возможность создавать ИИ-дикторов, имея их 30-секундное согласие на запись, при сохранении фонограммы с записью клонированного голоса дает выбор: сохранить ли файл с метаданными (где будет содержаться информация, что сервис Descript использовался при создании файла) или без них.

По нашему мнению, простым в реализации механизмом для маркирования генеративного контента, в первую очередь для его детекции среди аутентичного, является встраивание в метаданные файла информации о том, с использованием какого сервиса генерация была осуществлена. Дополнительным механизмом аутентификации может послужить также, например, включение идентификационного номера этой генерации. Но важно учесть технологическую возможность изменения или удаления информации в метаданных сгенерированных объектов, что существенно уменьшает эффективность использования подобной маркировки в целях судебного установления обстоятельств.

Некоторые механизмы правового регулирования могут быть взяты, например, из сферы применения блокчейн-технологии. Так, например, специализированные ресурсы [25] прогнозируют технологические решения, которые, по нашему мнению, могут быть имплементированы в отечественное законодательство.

Несмотря на очевидный потенциал блокчейн-технологии в маркировании контента, ее существенным недостатком остаются высокие технологические издержки на реализацию этого процесса.

Технологически водяные знаки подразделяются на два основных типа: хрупкие и прочные. Хрупкие легко разрушаются при манипуляциях пользователей с носителем, прочные – более устойчивы к манипуляциям, но их сложнее обнаружить обычному пользователю, не обладающему специальными знаниями.

Например, в качестве механизма проверки видео на предмет аутентичности может быть использована технология кодирования видео в неизменяемые блокчейны (в индустрии подобные технологии реализуют такие компании, как FactomAxiom).

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.

Таким образом, существующие методы маркирования контента представляют собой разнообразные технологии, каждая из которых имеет свои особенности и области применения. На данный момент можно выделить следующие основные виды маркирования:

– водяные знаки;

– встраивание служебной информации в метаданные файла;

– блокчейн-технологии для маркирования контента;

– визуальная маркировка.

Каждый из этих видов обладает как достоинствами, так и ограничениями, и выбор конкретного способа маркирования зависит от задач и характера самого контента.

В некоторых странах уже внедрены определенные механизмы правового регулирования распространения генеративного контента, связанные с его обязательной маркировкой. Так, например, п. 133 Акта о регулировании искусственного интеллекта, разработанного и одобренного Европейским парламентом [26], содержит меры по маркированию генеративного контента:

– водяные знаки, идентификация метаданных, криптографические методы подтверждения происхождения и подлинности контента, методы протоколирования, отпечатки пальцев или другие методы, в зависимости от обстоятельств;

– такие методы и способы должны быть совместимыми, эффективными и надежными, насколько это технически возможно, с учетом имеющихся методов или их сочетания, таких как водяные знаки, идентификация метаданных, криптографические методы подтверждения происхождения и подлинности контента, методы регистрации, отпечатки пальцев или другие методы, в зависимости от обстоятельств.

Конкретные меры правового регулирования оборота генеративного контента уже разработаны в Китайской Народной Республике, где на уровне межведомственного взаимодействия разработано Положение об администрировании службы глубокого синтеза информации в сети Интернет [27]:

– ст. 7 предусматривает ответственность компаний, предоставляющих ресурсы для синтеза контента;

– ст. 9 предписывает использование механизмов идентификации пользователей систем генеративного синтеза;

– ст. 16 обязывает производить маркировку генеративногоконтента. Подобная мера не только снижает криминогенный потенциал генеративного контента, но и создает условия для установления единого источника происхождения распространяемых файлов.

Прогрессивный опыт китайских коллег должен быть имплементирован в отечественное законодательство, но для успешного противодействия преступности, использующей генеративный контент для совершения преступлений, требуется в первую очередь разработка рамочных международных актов.

В США представлен законопроект «AI Disclosure Act of 2023» (Закон о раскрытии информации об искусственном интеллекте) [28], согласно которому предполагается наложить на пользователей обязанность раскрывать информацию о том, что что-либо было создано при помощи искусственного интеллекта.

Тем не менее важно отметить, что уже существуют сервисы по противодействию маркировке. Например, популярный сервис с открытым доступом достаточно успешно удаляет водяные знаки с различных изображений (http://dewatermark.ai.ru).

При этом отсутствие международных технологических стандартов для маркировки генеративного контента фактически сводит на нет любые меры правового регулирования. Без совместных координированных усилий по созданию единых механизмов правового регулирования маркировки генеративного контента все вышеперечисленные способы являются не эффективными.

Эксперты группы при ООН (Консультативный орган высокого уровня по искусственному интеллекту) предложили рекомендации по регулированию искусственного интеллекта в сентябре 2024 г. в докладе «Управление искусственным интеллектом в интересах человечества» [29]. В докладе отмечается опасность распространения дипфейков и подчеркивается важность разработки общих стандартов аутентификации контента и его цифрового происхождения.

О важности укрепления международного сотрудничества в области использования технологий искусственного интеллекта как одной из основных задач развития искусственного интеллекта в Российской Федерации говорится и в Национальной стратегии развития искусственного интеллекта.

Выводы

Принимая во внимание рассмотренную структуру субъектов, механизмы правового регулирования генеративного контента сводятся к следующему:

– установление обязанности технологических компаний и пользователей, создающих генеративный контент, использовать водяные знаки для маркирования контента в целях информирования иных лиц о генеративной природе контента;

– установление ответственности за отказ от маркировки или удаление какого-либо вида маркировки;

– установление ответственности за неправомерное использование биометрических персональных данных для создания дипфейка;

– учет общественной опасности деяний, совершаемых с использованием дипфейка в качестве средства совершения противоправного деяния;

– разработка и процессуальная регламентация средств обеспечения аутентичности цифровых доказательств в судопроизводстве.

Подводя промежуточный итог в рассмотрении вопроса о разработке механизмов правового регулирования и алгоритмах маркировки генеративного контента, следует заключить, что технологии искусственного интеллекта для создания генеративного контента уже получили обширное распространение на территории различных стран и правовых систем, стали частью многих видов человеческой деятельности. Указанное является причиной фактической невозможности установления какого-либо эффективного правового запрета на создание генеративного контента. Поэтому создание национальных правовых норм имеет важное значение, но не обеспечит информационную безопасность государства в долгосрочной перспективе. Только международное сотрудничество в сфере правого регулирования искусственного интеллекта и создание единых универсальных норм об обязательном маркировании генеративного контента будут действенным средством в борьбе с дипфейками и защите общества и государства от их пагубного воздействия.

СПИСОК ИСТОЧНИКОВ

1. Бодров Н. Ф., Лебедева А. К. Понятие дипфейка в российском праве, классификация дипфейков и вопросы их правового регулирования // Юридические исследования. 2023. № 11. С. 26–41.
2. Финансовый работник выплатил 25 миллионов долларов после видеозвонка с подставным финансовым директором. URL: https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html (дата обращения: 20.09.2024).
3. Admitad отказал в выплате 600к, так как сказал, что я дипфейк. URL: https://pikabu.ru/story/admitad_otkazal_v_vyiplate_600k_tk_skazal_chto_ya_dipfeyk_11295845?utm_source=linkshare&utm_medium=sharing (дата обращения: 20.09.2024).
4. Дипфейки приходят в финансовый сектор. URL: https://www.wsj.com/articles/deepfakes-are-coming-for-the-financial-sector-0c72d1e5 (дата обращения: 20.09.2024).
5. В правительстве не поддержали законопроект об уголовной ответственности за дипфейки. URL: https://tass.ru/obschestvo/17922853 (дата обращения: 24.09.2024).
6. О внесении изменений в часть первую Гражданского кодекса Российской Федерации : проект федерального закона № 718834-8 (внесен 16.09.2024 сенаторами Российской Федерации А. А. Клишасом, А. Г. Шейкиным, Н. С. Кувшиновой, Р. В. Смашневым, депутатом Государственной Думы Д. В. Бессарабовым). URL: https://sozd.duma.gov.ru/bill/718834-8 (дата обращения: 24.09.2024).
7. О внесении изменений в Уголовный кодекс Российской Федерации : проект федерального закона № 718538-8 (внесен 16.09.2024 депутатом Государственной Думы Я. Е. Ниловым, Сенатором Российской Федерации А. К. Пушковым). URL: https://sozd.duma.gov.ru/bill/718538-8 (дата обращения: 24.09.2024).
8. Бодров Н. Ф., Лебедева А. К. Понятие дипфейка (deepfake) в российском праве, его классификация и проблемы правового регулирования // Юридический вестник Дагестанского государственного университета. 2023. Т. 48, № 4 (68). С. 173–181.
9. Бодров Н. Ф., Лебедева А. К. Угрозы и вызовы в эпоху генеративного искусственного интеллекта с учетом криминогенного потенциала дипфейков // Санкт-Петербургский международный криминалистический форум : материалы междунар. науч.-практ. конф. Санкт-Петербург, 10–11 июня 2024 г. / сост. : А. Р. Акиев, Т. А. Бадзарадзе. СПб., 2024. С. 62–65.
10. Официальный отзыв от 22 июля 2024 г. № ДГ-П4-23438 на проект федерального закона «О внесении изменений в Уголовный кодекс Российской Федерации», вносимый в Государственную Думу депутатом Государственной Думы Я. Е. Ниловым. URL: https://sozd.duma.gov.ru/bill/594966-8?ysclid=m3cngyb18t186499919 (дата обращения: 24.09.2024).
11. Предупреждение о том, что изображения «мелких подделок» – это «следующее крупное мошенничество», которое поразит Британию. URL: https://www.dailymail.co.uk/news/article-13373513/shallowflake-scam-warning-car-insurance.html (дата обращения: 18.09.2024).
12. Бодров Н. Ф., Лебедева А. К. Дипфейк как объект судебной экспертизы // Национальные и международные тенденции и перспективы развития судебной экспертизы : сб. докладов науч.-практ. конф. с междунар. участием, г. Нижний Новгород, 22–23 мая 2024 г. Н. Новгород, 2024. С. 42–50.
13. EDIC-mini – диктофоны для вашей безопасности. URL: https://www.telesys.ru/Products/EM (дата обращения: 16.09.2024).
14. Ethics. URL: https://hourone.ai/ethics/ (дата обращения: 16.09.2024).
15. Представляем значок официальных учетных данных для контента. URL: https://c2pa.org/post/contentcredentials/ (дата обращения: 16.09.2024).
16. Content credential. Проверяйте содержимое, чтобы найти больше информации. URL: https://contentcredentials.org/verify (дата обращения: 16.09.2024).
17. Водяные знаки в режиме предварительного просмотра в службе Azure OpenAI. URL: https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/watermarks-in-preview-in-azure-openai-service/ba-p/4253344 (дата обращения: 26.09.2024).
18. Речь искусственного интеллекта Azure. URL: https://azure.microsoft.com/en-us/products/ai-services/ai-speech?msockid=04b5abc717656a530541bfde16f56b4b (дата обращения: 16.09.2024).
19. Идентификация контента, созданного искусственным интеллектом, с помощью SynthID. URL: https://deepmind.google/technologies/synthid/ (дата обращения: 26.09.2024).
20. Нейронный водяной знак Resemble может предотвратить использование ваших данных моделями ИИ, отслеживая их происхождение. URL: https://www.resemble.ai/watermarker/ (дата обращения: 20.09.2024).
21. Клонируйте свой голос, чтобы петь, говорить и не только... URL: https://myvocal.ai/billing (дата обращения: 20.09.2024).

22. YouTube Blog – Official Blog for Latest YouTube News&Insights. URL: https://blog.youtube (дата обращения: 20.09.2024).
23. Cloud Level. Ваши облака попадают сюда. URL: https://about.fb.com/news/ (дата обращения: 20.09.2024).
24. Proactive Detection of Voice Cloning with Localized Watermarking / Robin San Roman, Pierre Fernandez, Alexandre Défossez, Teddy Furon, Tuan Tran, Hady Elsahar. URL: https://arxiv.org/abs/2401.17264 (дата обращения: 20.09.2024).
25. Deep Fake Challenge. Любая идентификация имеет нулевую ценность, если ее можно обмануть. URL https://deepfakechallenge.com/ (дата обращения: 20.09.2024).
26. European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)). URL: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html (дата обращения: 06.09.2024).
27. Государственное информационное управление Интернета и три других ведомства выпустили «Положение об управлении углубленным обобщением информационных услуг Интернета». URL: https://www.cac.gov.cn/2022-12/11/c_1672221949318230.htm (дата обращения: 06.09.2024).
28. H.R.3831 – AI Disclosure Act of 2023. URL: https://www.congress.gov/bill/118th-congress/house-bill/3831/text?s=1&r=1 (дата обращения: 06.09.2024).
29. Управление искусственным интеллектом на благо человечества : заключительный доклад. URL: https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_ru.pdf (дата обращения: 30.09.2024).

REFERENCES

1. Bodrov N.F., Lebedeva A.K. The concept of deepfake in Russian law, classification of deepfake and issues of their legal regulation. *Yuridicheskie issledovaniya = Legal Studies,* 2023, no. 11, pp. 26–41. (In Russ.).
2. Finance worker pays out $25 million after video call with deepfake chief financial officer. *CNN.* Available at: https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html (accessed September 20, 2024).
3. *Admitad otkazal v vyplate 600k, t.k skazal chto, ya dipfeik* [Admitad refused to pay 600k, because it said that I was a deepfake]. Available at: https://pikabu.ru/story/admitad_otkazal_v_vyiplate_600k_tk_skazal_chto_ya_dipfeyk_11295845?utm_source=linkshare&utm_medium=sharing (In Russ.). (Accessed September 20, 2024).
4. Deepfakes are coming for the financial sector. *The Wall Street Journal*. Available at: https://www.wsj.com/articles/deepfakes-are-coming-for-the-financial-sector-0c72d1e5 (accessed September 20, 2024).
5. The government did not support a draft law on criminal liability for deepfakes. *Informatsionnoe agentstvo TASS* [TASS News Agency]. Available at: https://tass.ru/obschestvo/17922853 (In Russ.). (Accessed September 24, 2024).
6. *O vnesenii izmenenii v chast' pervuyu Grazhdanskogo kodeksa Rossiiskoi Federatsii : proekt federal'nogo zakona No. 718834-8 (vnesen 16.09.2024 senatorami Rossiiskoi Federatsii A.A. Klishasom, A.G. Sheikinym, N.S. Kuvshinovoi, R.V. Smashnevym, deputatom Gosudarstvennoi Dumy D.V. Bessarabovym)* [On Introducing Amendments to Part One of the Civil Code of the Russian Federation: Draft Federal Law No. 718834-8 (introduced on September 16, 2024 by Senators of the Russian Federation A.A. Klishas, A.G. Sheikin, N.S. Kuvshinova, R.V. Smashnev, State Duma Deputy D.V. Bessarabov)]. Available at: https://sozd.duma.gov.ru/bill/718834-8 (accessed September 24, 2024).
7. *O vnesenii izmenenii v Ugolovnyi kodeks Rossiiskoi Federatsii: proekt federal'nogo zakona № 718538-8 (vnesen 16.09.2024 deputatom Gosudarstvennoi Dumy Ya.E. Nilovym, Senatorom Rossiiskoi Federatsii A.K. Pushkovym)* [On Amendments to the Criminal Code of the Russian Federation: Draft Federal Law No. 718538-8 (introduced on September 16, 2024 by State Duma Deputy Ya.E. Nilov, Senator of the Russian Federation A.K. Pushkov)]. Available at: https://sozd.duma.gov.ru/bill/718538-8 (accessed September 24, 2024).
8. Bodrov N.F., Lebedeva A.K. The concept of a deepfake in Russian law, its classification of deepfakes and problems of legal regulation. *Yuridicheskii vestnik Dagestanskogo gosudarstvennogo universiteta = Herald of Dagestan State University*, 2023, vol. 48, no. 4 (68), pp. 173–181. (In Russ.).
9. Bodrov N.F., Lebedeva A.K. Threats and challenges in the era of generative artificial intelligence, taking into account the criminogenic potential of deepfakes. In: *Sankt-Peterburgskii mezhdunarodnyi kriminalisticheskii forum: materialy mezhdunar. nauch.-prakt. konf.* [Saint Petersburg International Forensic Forum: proceedings of the international scientific and practical conference. Saint Petersburg, June 10–11, 2024]. Saint Petersburg, 2024. Pp. 62–65. (In Russ.).
10. *Ofitsial'nyi otzyv ot 22 iyulya 2024 g. No DG-P4-23438 na proekt federal'nogo zakona "O vnesenii izmenenii v Ugolovnyi kodeks Rossiiskoi Federatsii", vnosimyi v Gosudarstvennuyu Dumu deputatom Gosudarstvennoi Dumy Ya.E. Nilovym* [Official review No. DG-P4-23438 of July 22, 2024 on the draft federal law "On Amendments to the Criminal Code of the Russian Federation", submitted to the State Duma by State Duma Deputy Ya.E. Nilov]. Available at: https://sozd.duma.gov.ru/ (accessed September 24, 2024).
11. Warning that "shallowfake" images are the "next big scam" to hit Britain: Fraudsters are mocking-up pictures of car damage to con insurers – with number of cases surging by 300 per cent in a year. *Mail Online*. Available at: https://www.dailymail.co.uk/news/article-13373513/shallowflake-scam-warning-car-insurance.html (accessed September 18, 2024).
12. Bodrov N.F., Lebedeva A.K. Deepfake as an object of forensic examination. In: *Natsional'nye i mezhdunarodnye tendentsii i perspektivy razvitiya sudebnoi ekspertizy: sb. dokladov nauch.-prakt. konf. s mezhd. uchastiem, g. Nizhnii Novgorod, 22–23 maya 2024 g*. [National and international trends and prospects for the development of forensic examination: collection of reports of the scientific and practical conference with international participation, Nizhny Novgorod, May 22–23, 2024]. Nizhny Novgorod, 2024. Pp. 42–50. (In Russ.).
13. *EDIC-mini — diktofony dlya vashei bezopasnosti* [EDIC-mini voice recorders for your safety]. Available at: https://www.telesys.ru/Products/EM (accessed September 16, 2024).
14. *Ethics*. Available at: https://hourone.ai/ethics/ (accessed September 16, 2024).
15. *Introducing official content credentials icon.* Available at: https://c2pa.org/post/contentcredentials/ (accessed September 16, 2024).

16. *Content credentials*. Available at: https://contentcredentials.org/verify (accessed September 16, 2024).
17 *Watermarks in preview in Azure OpenAI Service*. Available at: https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/watermarks-in-preview-in-azure-openai-service/ba-p/4253344 (accessed September 26, 2024).
18. *Azure AI Speech*. Available at: https://azure.microsoft.com/en-us/products/ai-services/ai-speech?msockid=04b5abc717656a530541bfde16f56b4b (accessed September 16, 2024).
19. *DeepMind.* Available at: https://deepmind.google/technologies/synthid/ (accessed September 26, 2024).
20. *Resemble.ai.* Available at: https://www.resemble.ai/watermarker/ (accessed September 20, 2024).
21. *Myvocal.ai.* Available at: https://myvocal.ai/billing (accessed September 20, 2024).
22. YouTube Blog – Official Blog for Latest YouTube News&Insights. Available at: https://blog.youtube. (accessed September 20, 2024).
23. *Cloud Level.* Available at: https://about.fb.com/news/ (accessed September 20, 2024).
24. Roman R.S., Fernandez P., Défossez A., Furon T. et al. *Proactive detection of voice cloning with localized watermarking*. Available at: https://arxiv.org/abs/2401.17264 (accessed September 20, 2024).
25. *Deep fake challenge*. Available at: https://deepfakechallenge.com/ (accessed September 20, 2024).
26. *European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))*. Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html (accessed September 6, 2024).
27. [Departmental rule "Regulation on the Administration of the Deep Information Synthesis Service on the Internet"]. Available at: https://www.cac.gov.cn/2022-12/11/c_1672221949318230.htm (accessed September 6, 2024).
28. *H.R.3831 – AI Disclosure Act of 2023*. Available at: https://www.congress.gov/bill/118th-congress/house-bill/3831/text?s=1&r=1 (accessed September 6, 2024).
29. *Governing AI for humanity: final report.* Available at: https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_ru.pdf (accessed September 30, 2024).

СВЕДЕНИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

НИКОЛАЙ ФИЛИППОВИЧ БОДРОВ – кандидат юридических наук, президент Международной общественной организации «Союз криминалистов и криминологов», доцент кафедры судебных экспертиз Университета имени О. Е. Кутафина (МГЮА), Москва, Россия, bodrovnf@gmail.com, https://orcid.org/0000-0002-9005-3821

АНТОНИНА КОНСТАНТИНОВНА ЛЕБЕДЕВА – кандидат юридических наук, член Международной общественной организации «Союз криминалистов и криминологов», доцент кафедры судебных экспертиз Университета имени О. Е. Кутафина (МГЮА), Москва, Россия, tonya109@yandex.ru, https://orcid.org/0009-0004-9344-2103

NIKOLAI F. BODROV – Candidate of Sciences (Law), President of the Union of Criminalists and Criminologists, Moscow, Russia, associate professor at the Forensic Expertise Department of the Kutafin Moscow State Law University (MSAL), Moscow, Russia, bodrovnf@gmail.com, https://orcid.org/0000-0002-9005-3821

ANTONINA K. LEBEDEVA – Candidate of Sciences (Law), Member of the Union of Criminalists and Criminologists, Moscow, Russia, associate professor at the Forensic Expertise Department of the Kutafin Moscow State Law University (MSAL), Moscow, Russia, tonya109@yandex.ru, ttps://orcid.org/0009-0004-9344-2103

Статья поступила 10.10.2024