DOI 10.36622/1682-7813.2025.28.1.009
УДК 004: 327:355.014

О ПОДХОДАХ В РАСПОЗНАВАНИИ ДИПФЕЙКОВ И СИНТЕТИЧЕСКИХ МЕДИА

С.А. Воронов, А.Д. Косолапов, А.В. Скробач

Благодаря высокой степени информатизации всех сфер деятельности, для осуществления информационно-психологического воздействия могут применяться различные технологии, опирающиеся на всю информационную инфраструктуру. Информация, как предмет информационного обмена, играет важную роль и позволяет менять поведение людей. Целью работы является анализ существующих методов распознавания фейковой информации (дипфейков), генерируемой при помощи технологий искусственного интеллекта. В работе авторы указали на свою позицию по различию между терминами «дипфейк» и «синтетическое медиа», которые зачастую ряд исследователей синонимизируют. Приводится классификация дипфейков по типам генерации и по целям. В результате анализа авторы выделяют три основных подхода к распознаванию дипфейков: поиск артефактов, использование технологии блокчейн и использование мастер-данных. Подробнее раскрыт подход, который, по мнению авторов, является наиболее интересным, основанный на управлении мастер-данными (MDM). Проведенный анализ позволяет структурировать информацию о применяемых технологиях в распознавании провокационной или ложной информации, генерируемой с помощью искусственного интеллекта, в частности, алгоритма generative adversarial netwok (GAN).

Ключевые слова: дипфейк, синтетические медиа, распознавание дипфейков, искусственный интеллект, информационное противоборство, инструменты информационно-психологического воздействия

Введение

Информация как объект отношений всегда играла высокую роль. С ростом видов и способов распространения информации изменилась и сущность ее. Зависимость от информации, от средств доставки ее до потребителя, важности, делает участников информационного обмена не всегда восприимчивым к ее объективности. Скорость принятия решений, сроки обработки и объемы информации не оставляют порой времени для анализа информации, и вопрос правдивости или ложности становиться не первостепенным. Именно такие факторы позволяют использовать недостоверную информацию для решения задач, направленных на негативное воздействие. Благодаря высокой степени информатизации всех сфер деятельности, для осуществления информационно-психологического воздействия могут применяться различные технологии, опирающиеся на всю информационную инфраструктуру.

СМИ как основной источник информации перестал быть единственным каналом распространения информации. Современные технологии и, зачастую подконтрольность тех или иных СМИ определенным влиянию, стали триггером для развития блогосферы, и теперь каждый, кто находиться ближе к месту события или первым опубликовал новость, может быть источником информации. Профессиональная журналистика сосуществует параллельно с работой блогеров, их частным мнением, а федеральные или местные СМИ соревнуются с социальными сетями и метаплатформами. Именно материалы от обычных пользователей, распространяемые через различные социальные медиа, в большей степени могут быть непроверенными, являться слухами, выдумкой или заведомо содержать ложную информацию. В профессиональной журналистике фальшивую информацию называют фейком, от английского слова fake (вымышленный, лукавый, фальшивый, шутка и т. д.).

Обострение отношений на мировой арене стало очередным толчком для развития фейковой информации. Скачек объема такой информации однозначно обозначил цели ее распространения, что заставило вносить необходимые поправки в действующее законодательство в части противодействия недостоверной информации [1].

По данным исследования [2], с 2022 г. число фейков выросло в 6 раз, а интерес аудитории вырос по сравнению с предыдущим годов на 68%. Основным же источником распространения фейковой информации стали социальные сети и мессенджеры мгновенного обмена информацией (ВК (24%), Одноклассники (20%), Телеграмм (20%)), а на СМИ пришлось лишь 12% фейковых новостей.

О синтетических медиа и дипфейках, сравнение и классификация

Развитие технологий искусственного интеллекта (AI) в последние годы не только позволило вывести рутинную обработку больших объемов данных на новый уровень, но и создало определенные угрозы информационной безопасности как для людей, так и для отдельных направлений бизнеса и государства. Текстовый, аудио и видео материал, создаваемый с помощью технологий искусственного интеллекта (ИИ), или другими словами, синтетические медиа, от невинных форм развлечения все чаще стали применяться для совершения противоправных действий или задач по дискредитации отдельных личностей, категорий граждан и целых государств.

Доступность сервисов с моделями машинного обучения (ML-модели) позволяет массовому пользователю создавать контент без профессиональных навыков видеомонтажа, программирования, редактирования изображений.

Сам процесс замены фрагмента изображения с помощью программ не новый, и по названию одного из самых распространенных графических редакторов зачастую использовался термин «фотошоп» (отфотошопленное изображение и т. д.). Процесс покадрового изменения видеофрагмента таким методом вызывал существенные сложности и временные затраты. Развитие киноиндустрии при работе со спецэффектами и компьютерному изменению возраста актеров стало по мнению ряда экспертов предтечей появлению дипфейка. Однако сам термин «дипфейк» появился в 2017 году, когда пользователь с ником Deepfakes на платформе Reddit разместил несколько фейковых видео, где лица актрис из порнофильмов заменены на лица известных актрис (Галь Гадот, Тейлор Свифт, Скарлетт Йоханссон и др.).

Само слово «дипфейк» происходит от двух терминов: deep learning, глубокий метод машинного обучения на нейросетях, и fake – подделка.

Другой известный эпизод, который создан с помощью программ FakeApp и Adobe After Effects, связан с публикацией в 2018 г. фальшивого видео, где якобы экс-президент США Барак Обама оскорбляет действующего тогда президента Дональда Трампа. В дальнейшем было опубликовано много различных дипфек-видео в отношении известных личностей и представителей власти. Эти события показали серьезность возникшей новой угрозы и в 2019 г. появился первый нормативно-правовой акт AB 730 [3] в штате Калифорния, США, ограничивающий применение искусственного интеллекта, в частности, запрещающий распространение дипфейков в период предвыборной кампании в 2020 г.

Помимо негативных примеров применения дипфейков есть и вполне безобидные, это в кино и рекламе. В рекламе от Сбербанка восстановили образ Жоржа Милославского из кинофильма «Иван Васильевич меняет профессию», а в рекламе Citroen для продвижения бренда использовали цифровой образ Джона Леннона. С использованием в рекламе персонажей, которых нет уже в живых, вопрос не совсем однозначный в части уместности того или иного образа рядом в рекламируемым товаром, но тут все дело в получении материальной выгоды соответствующими родственниками, для которых этический вопрос не всегда стоит на первом месте. В киноиндустрии же порой сами поклонники ратуют «воскрешению» того или иного героя. Для киноэпопеи

«Звездные войны», эпизод 9, продюсеры воссоздали умершую актрису Кэрри Фишер (1956–2016 гг.), а для «Изгой-один: Звёздные войны» воскресили персонажа, которого играл актер Питер Кушинг, скончавшегося в 1994 году. В обоих случаях, использование цифрового образа умерших людей, а другими словами, дипфейка, санкционировано ближайшими родственниками и не несет ущерба его обладателю.

Первоначально термин «дипфейк» употребляли к методу замены лица человека на видеофрагменте, однако со временем это распространилось на все поддельные материалы: аудио, видео, текст, изображения, где применялся механизм генерации синтетического медиа.

Технология Deepfake «представляет собой методику компьютерного синтеза изображения, основанную на искусственном интеллекте, которая используется для соединения и наложения существующих изображений и видео на исходное изображение или видеоролик» [4].

Более полным определением видеться следующее – «синтез правдоподобных поддельных изображений, видео и звука при помощи искусственного интеллекта. Также дипфейками называют контент, полученный в результате этого синтеза» [5].

В отличие от техноцентрического подхода, в основе которого лежат применяемые методы и технологии для создания дипфейка, нормативный подход опирается на возможный вред, который может быть причинен обладателю образа или другим участникам информационного обмена. Другими словами, дипфейк рассматривается как незаконный акт использования чужого голоса или лица без согласия его обладателя для совершения деяния, направленного на дезинформацию других лиц.

Наибольшую опасность дипфейки представляют при мошеннических действиях. Одним из крупных известных таких примеров служит случай в ОАЭ, когда банковский служащий перевел 35 млн. долларов на неизвестный счет. Также известны множество случаев, когда блогеры или мошенники применяли технологии генерации дипфейка в режиме видеоконференции при общении с представителями власти для получения конфиденциальной информации, которая становилась достоянием общественности, или для создания условий для провокаций.

Вне зависимости от используемых подходов в определении дипфейка, стоит отметить, что зачастую, ряд экспертов и исследователей все синтетические медиа отождествляют с дипфейками. Мы придерживаемся позиции, что синонимизация терминов «дипфейк» и «синтетическое медиа» ошибочна, и основное различие заключается в целях создания такого контента и морально-нормативной релятивности понятия дипфейк [6]. Иначе говоря, лишь скрытая цель (угроза нанесения ущерба человеку, государству и т.п.) или неправомерное использование образа (без согласия правообладателя) в сгенерированном синтетическом контенте характеризует дипфейк.

Несмотря на то, что синтетические медиа, созданные для развития киноиндустрии и музыки на первый взгляд не несут явных угроз, а предназначены, в первую очередь, для изменения возраста актера, «воскрешения» известных личностей, улучшение качества дубляжа, уменьшения затрат на производство конечного продукта, но это не совсем так. В данном случае, использование современных технологий также может наносить неприемлемый ущерб актерам, чей образ (внешний вид, голос), единожды был оцифрованн, и в дальнейшем может использоваться киностудией. В том числе по этой причине в 2023 году Американская федерация артистов телевидения и радио (SAG-AFTRA) устроила забастовки актеров, требовавших, в частности, ограничить использование в киноиндустрии технологий искусственного интеллект. Что привело в итоге к соглашению между сторонами, и в Калифорнии запретили использование цифровых копий образов и голоса актеров без соответствующих разрешений.

Доступность технологий искусственного интеллекта не только упростило ряд операций по работе с графикой, но и создало новые угрозы применения результатов деятельности

ИИ в противоправных или провокационных целях.

Если технологии обработки статичной 2D графики с помощью графических редакторов давно известны и не вызывают ни у кого удивления, то сейчас мейнстрим – технологии обработки видео и аудио потоков в режиме реального времени. С уровнем развития технологий машинного обучения растет и качество дипфейков.

Таким образом, в целях генерации дипфеков (deepfake) можно выделить следующие:

- развлекательные. Создание развлекательного контента на платформах и ресурсах. Тонкая грань между созданием контента развлекательного характера и высмеиванием, булингом, ставит такого рода дипфейки рядом с деструктивными [6, 7];

- коммуникационные. Генерация образов для представительских мероприятий и при обучении, дополнение формы визуальной коммуникации человека с «роботом» в более привычном виде; привлечение внимания к проблеме, событию. В перспективе, генерация таких образов станет неотъемлемой частью сопровождения представительских мероприятий, в образовательной деятельности для сопровождения курса самообучения, анимирования голоса и создания иллюзии присутствия преподавателя, в туризме — в роли гида иди экскурсовода, в отельном бизнесе — ресепшен и т.д. [8];

- мошеннические. Противоправная деятельность с целью получения финансовой или экономической выгоды [9];

- деструктивные. Нанесение имиджевого ущерба отдельным личностям, месть, информационное противоборство, дискредитация органов власти, политическая борьба [10, 11].

Среди визуальных дипфейков выделяют следующие типы:

- замена лица. От простых идей автоэнкодеров до более продвинутых, к примеру таких, как использование двух моделей: где первая модель, Adaptive Embedding Integration Network, используется для осуществления самого по себе переноса лица, а вторая, Heuristic Error Acknowledging Network, используется для улучшения качества итогового переноса;

- перенос выражения лица или манипуляция мимикой;

- синхронизация губ в ходе разговора;

- добавление атрибутов, косметическое изменение лица (очки, затемнение или осветление кожи, изменение прически и т.д.);

- генерация несуществующего образа лица [12, 13].

Среди аудиальных дипфейков выделяют два типа:

1) синтез текста в речь по аудиофрагменту голоса;

2) преобразование голоса в режиме реального времени [13].

О подходах в распознавании синтетических медиа и дипфейков

В основе дипфейка в большинстве случаев лежат технологии глубокого обучения, в частности, генеративно-состязательные сети (GAN), которые включают две глубоких нейронных сети, «соревнующихся» между собой: генератор и дискриминатор.

Первая сеть генерирует изображения, похожие на целевые, а вторая, обученная на наборе данных, состоящем из реальных фотографий, пытается отличить «правильные» образцы от «неправильных», оценивая таким образом достоверность сгенерированных данных.

Если для создания дипфейков

используются модели машинного обучения, то и для распознавания могут использоваться аналогичные приемы. Такой подход будет основан на анализе исходного изображения, наличия в нем несогласованных элементов, к примеру расхождение при монтаже видеоряда мимики лица и наложенной речи, неестественное движение глаз и моргание [14].

Нейросети ищут низкоуровневые артефакты на изображениях, которые слаборазличимы человеческим глазом: корректность теней и освещения, качество изображения различных деталей на переднем и заднем плане и т. д. На видеозаписях сравнивается движение губ и произносимого ими звука, и такой способ дает точность от 73,5 до 97,6%. Анализу также подвергаться могут изменение цвета кожи, размеры дефектов кожного покрова (родинки, шрамы).

Другой подход предполагает использовать технологию блокчейн, для фиксации официального контента [15]. Так, объединение ряда корпораций, среди которых есть Microsoft, Intel, Adobe, Sony, Nikon, Twitter, планируют разработать и внедрить единый стандарт, позволяющий подтверждать уникальность фото и видео. В основе планируемого стандарта лежат блокчейн-технологии, благодаря которым будет проверяться подлинность материалов. Если рассматривать вопрос защиты видеоконференции от дипфейка, то одним из решений может быть встраивание в аудио- или видеозапись сигнала, который нельзя подделать, что позволит отличить в режиме реального времени реальные аудио и видео от поддельных.

Третий подход предполагает использовать мультимодальную модель, использующую граф знаний. Речь о семантических данных или атрибутах, которые можно проверить на достоверность на основе открытой информации в сети [16].

Такой подход предлагает сопоставлять факты из различных информационных систем, в основе которого лежит управление данными (MDM), используемое для распознавания дипфейков. Мастер-данные — это консолидированные данные, или данные, хранящиеся в различных информационных системах, а процесс их консолидации — управлением мастер-данными (MDM). Пример использования сразу двух подходов, аналог блокчейн и MDM, это проверка фотографии из социальной сети, на которой несколько отметок разных пользователей (себя или место), указывающее на подлинность фотографии. Такой подход более адаптирован для распознавания текстового контента, содержащего фейки. Одно из распространенных направлений проверки фактов имеет название fact checking.

Рассмотрим подход, который, по нашему мнению, считается наиболее перспективным, предполагающий анализ информации из различных источников. Сбор и первичная обработка информации поддается алгоритмизации и вполне по силам существующим технологиям искусственного интеллекта. Исходя из этого можно выделить следующие задачи:

- поиск альтернативных источников данных — новостные ленты, социальные сети, фотобанки;
- использование различных алгоритмов сопоставления данных [16], для того чтобы определить контекст альтернативного потока данных и факта. Для этого необходимо проанализировать метаданные объекта. Процесс MDM как раз подразумевает создание и согласование единых метаданных для консолидации информации;
- проверка данных — при выявлении фактов, указывающих или описывающих одно и тоже событие, происходит сравнение кластеризованных данных.

Однако при таком подходе при анализе аудиоданных проблематично оценить эмоциональную составляющую, которая зачастую может иметь существенное значение. Также нельзя однозначно утверждать, что проверяемость данных равнозначна достоверности, особенно при сравнении данных из официальных источников и постов в социальных сетях или телеграмм-каналов (блогов). MDM-подход имеет слабые стороны при «горячих

звонках», когда сейчас поступает видеозвонок или в чате аудиосообщение, и необходимо реагировать в режиме реального времени, т. е. необходимо некоторое время и вычислительные мощности для анализа контента.

Для распознавания интерактивного дипфейка (в режиме реального времени) могут помочь методы отклонения от стандартных моделей общения (разговора, поведения). Дипфейк — это, зачастую, результат синтетического преобразования исходного фрагмента контента по шаблону, поэтому отклонение от шаблона поможет выявить уязвимости такого воздействия. К примеру, при видеозвонке, просьба к абоненту покрутить камеру для оценки обстановки или изменить положение головы (лица), может нарушить существующий алгоритм и заложенные шаблоны. При анализе дипфейков без обратной связи такой подход уже нельзя реализовать.

Таким образом, необходимо использовать разные подходы к анализу и распознаванию дипфейков исходя их вида и применяемых технологий создания синтетических медиа.

Возвращаясь к MDM-подходу, по мнению Кузнецова С.В., «для эффективной борьбы с дипфейками следует построить замкнутый цикл обработки поступающей информации» [17]:

- коррекция алгоритмов машинного обучения при сопоставлении контента на основе результатов обработки фейковой информации;
- наделение источников информации определенным уровнем доверия, основываясь на результатах анализа качества и достоверности данных, и который может динамически изменяться;
- подготовка данных анализа информации с понятными критериями, по которым тот или иной контент был отнесен с определенной степенью вероятности к фейку, для последующего принятия экспертного решения.

Заключение

Искусственный интеллект стал сильным инструментом в решении задач по генерации синтетического контента, и он также может послужить для обратного процесса распознавания. Востребованность к выявлению дипфеков будет расти пропорционально их появлению. Развитие технологий искусственного интеллекта и рост доступности соответствующего инструментария приведет к росту синтетической информации, и с постоянно увеличивающимся объемом такой информации человеку будет крайне трудно справиться без вспомогательных технологий. Необходимо создавать соответствующие системы противодействия и распознавания синтетической информации и определенного участия законодательного института в регулировании этих процессов.

Список литературы

1. Уголовная ответственность за публичное распространение заведомо ложной информации. // URL: http://www.kremlin.ru/events/president/news/67908 (дата обращения: 20.12.2025).

2. Количество фейков в сети выросло в 6 раз // Информационное агенство ТАСС. URL:http://tass.ru/obshestvo/16642301 (дата обращения: 20.12.2025).

3. Assembly Bill No. 730 URL: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730 (дата обращения: 20.12.2025).

4. Иванов В.Г. Deepfakes: Перспективы применения в политике и угрозы для личности и национальной безопасности / В.Г. Иванов, Я.Р. Игнатовский // Вестник российского университета дружбы народов. Серия: Государственное и муниципальное управление. 2020. №4. С. 379-386.

5. Дипфейк, deepfake // URL: https://encyclopedia.kaspersky.ru/glossary/deepfake/ (дата обращения 11.12.24 г.)

6. Воронин И.А. Дипфейки: Современное понимание, подходы к определению, характеристики, проблемы и перспективы / И.А. Воронин, Д.П. Гавра // Российская школа связей с общественностью. 2024. № 33. С.28-47.

7. Kshetri N. The economics of deepfakes //Computer. 2023. T. 56. N. 8. p. 89-94.

8. Толстых М.Ю. Средства противостояния фейковой информации как угрозе информационной и национальной безопасности // Криминологический журнал. 2023. № 3. С. 291-298

9. Bateman J. Deepfakes and synthetic media in the financial system: Assessing threat scenarios. – Carnegie Endowment for International Peace., 2022. p. 52.

10. Delfino R. A. Pornographic deepfakes: The case for federal criminalization of revenge porn's next tragic act //Fordham L. Rev. 2019. T. 88. p. 887.

11. Appel M., Prietzel F. The detection of political deepfakes //Journal of Computer-Mediated Communication. 2022. T. 27. No 4.

12. Паркин А. Дипфейки: как определить подделку и обезопасить пользователей от мошенников URL: https://www.secuteck.ru/articles/dipfejki-kak-opredelit-poddelku-i-obezopasit-polzovatelej-ot-moshennikov (дата обращения 18.12.24).

13. Masood M. et al. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward //Applied intelligence. 2023. T. 53. N. 4. p. 3974-4026.

14. Digvijay Yadav, Sakina Salmani. Deepfake: A Survey on Facial Forgery Technique Using Generative Adversarial Network. May 2019. Conference: 2019 International Conference on Intelligent Computing and Control Systems (ICCS) DOI:10.1109/ICCS45141.2019.9065881.

15. Блокчейн против дипфейков URL: https://rspectr.com/articles/blokchejn-protiv-dipfejkov?ysclid=m4ayyr2jwa159875614 (дата обращения 11.12.24).

16. Yuliang Li, Jinfeng Li, Yoshihiko Suhara, Jin Wang, Wataru Hirota, Wang-Chiew Tan. 2021. Deep Entity Matching: Challenges and Opportunities. J. Data and Information Quality 13, 1, Article 1 (March 2021), 17 p.

17. Kuznetsov S.V., Koznov D.V. Master data management in an iterative approach. Ontology of Designing. Vol 11, N 2, p. 170-184.

Военная ордена Жукова академия войск национальной гвардии,
г. Санкт-Петербург, Россия
Military Order of Zhukov Academy of the National Guard Troops,
St. Petersburg, Russia

Поступила в редакцию 22.12.2024

Сведения об авторах

Воронов Сергей Алексеевич – кандидат педагогических наук, военная ордена Жукова академия войск национальной гвардии, e-mail: voronov-sci@mail.ru

Косолапов Александр Дмитриевич – кандидат педагогических наук, доцент, военная ордена Жукова академия войск национальной гвардии, e-mail: a.kosolapov@mail.ru

Скробач Александр Владимирович – кандидат физико-математических наук, доцент, военная ордена Жукова академия войск национальной гвардии, e-mail: aleksandr-scrobach@yandex.ru

ON APPROACHES TO RECOGNITING DEEPFAKES AND SYNTHETIC MEDIA

S.A. Voronov, A.D. Kosolapov, A.V. Skrobach

Due to the high level of informatization of all spheres of activity, various technologies based on the entire information infrastructure can be used to implement information and psychological influence. Information, as a subject of information exchange, plays an important role and allows changing people's behavior. The purpose of the work is to analyze existing methods for recognizing fake information (deepfakes) generated using artificial intelligence technologies. In the work, the authors indicated their position on the difference between the terms "deepfake" and "synthetic media", which are often synonymized by a number of researchers. A classification of deepfakes by generation type and purpose is provided. As a result of the analysis, the authors identify three main approaches to recognizing deepfakes: searching for artifacts, using blockchain technology and using master data. The approach, which, in the authors' opinion, is the most interesting, based on master data management (MDM), is described in more detail. The conducted analysis allows us to structure information about the technologies used in recognizing provocative or false information generated using artificial intelligence, in particular the generative adversarial network (GAN) algorithm.

Keywords: fakes, synthetic media, fake recognition, artificial intelligence, information warfare, tools of information and psychological influence

Submitted 22.12.2024

Information about the authors

Sergey A. Voronov – Cand. Sc. (Pedagogical), Military Order of Zhukov Academy of the National Guard Troops, e-mail: voronov-sci@mail.ru

Alexander D. Kosolapov – Cand. Sc. (Pedagogical), Associate Professor, Military Order of Zhukov Academy of National Guard Troops, e-mail: a.kosolapov@mail.ru

Alexander V. Skrobach – Cand. Sc. (Physico-Mathematical), Associate Professor, Military Order of Zhukov Academy of the National Guard Troops, e-mail: aleksandr-scrobach@yandex.ru