Филология: научные исследования
Правильная ссылка на статью:
Неренц Д.В. Дипфейк как одна из главных информационных угроз XXI века // Филология: научные исследования. 2025. № 9. DOI: 10.7256/2454-0749.2025.9.75041 EDN: TQDNWC URL: https://nbpublish.com/library_read_article.php?id=75041

Дипфейк как одна из главных информационных угроз XXI века

Неренц Дарья Валерьевна

кандидат филологических наук

доцент, кафедра журналистики, Российский государственный гуманитарный университет

125993, Россия, Московская область, г. Москва, Миусская площадь, 6, ауд. 525

ya.newlevel@yandex.ru

Статья из рубрики "Фейки"

DOI:

10.7256/2454-0749.2025.9.75041

EDN:

TQDNWC

Дата направления статьи в редакцию:

01-07-2025

Аннотация: Искусственный интеллект (ИИ) – это инструмент, который способен предоставить массу возможностей тем, кто сумеет им эффективно воспользоваться. Однако, как всегда при глобальных изменениях в жизни общества, есть и обратная сторона медали – пользоваться им учатся не только во благо, но и для реализации собственных планов и идей, которые могут наносить серьезный ущерб и отдельному человеку, и населению в целом. Дипфейк как продукт, созданный ИИ, является на сегодняшний день одной из самых серьезных информационных угроз, поскольку способен обмануть не только неискушенного пользователя, но и профессионального работника сферы IT. Каждый день в медиаполе появляется все больше резонансных примеров дипфейков, которые невозможно отличить от реальной аудио- или видеозаписи. Цель данной статьи – представить характерные особенности дипфейка как цифрового продукта, представить типологию таких публикаций по способу создания и целевым установкам, а также описать возможные маркеры распознавания подделок в онлайн-среде. Методология исследования основана на системно-структурном анализе медийного пространства, позволяющим типологизировать дипфейки, а также на методе

анализа контента, благодаря которому удалось выделить цели и характерные черты различных дипфейков. В работе также применен метод описания и метод обобщения. В качестве эмпирической базы выступил снегенерированный контент фото-, аудио- и видеоформата, вызвавший резонанс в обществе и публично разоблаченный либо самими создателями, либо средствами массовой информации, либо внимательными пользователями. Хронологические рамки исследования – 2020–2025 годы. Всего было проанализировано 69 дипфейков. Материал был собран методом сплошной выборки и содержал в себе публикации, упомянутые в СМИ (на телеканалах «Первый канал», НТВ и «Россия 1», платформе «Смотрим», интернет-изданиях «Лента.ру» и «Газета.ру», сайтах информационных агентств «РИА Новости» и ТАСС). Подобный подход позволил сформировать представление о тематике, цели и методах тиражирования дипфейков, а также выделить наиболее эффективные способы их распознавания. Данные маркеры являются актуальными и эффективными на современном этапе, поскольку позволяют максимально оперативно и эффективно отличить генерацию, что особенно важно в условиях постоянного совершенствования механизмов функционирования нейросетей.

Ключевые слова:

дипфейк, информационная угроза, СМИ, журналистика, фейк, искусственный интеллект, медиапространство, нейросеть, технология, видео

Разговоры об искусственном интеллекте и его возможностях сегодня не просто не прекращаются, но многократно множатся. Его обсуждают ученые, практикующие специалисты, работники IT-сферы, государственные служащие. Уже очевидно, что мы находимся в некоем переломном моменте и становимся свидетелями появления совершенно новой модели занятости, где рядом с человеком будут работать цифровые сотрудники (так называемые интеллектуальные агенты) и роботизированные системы. Автоматизация рутинных процессов, о которых уже написано и сказано множество слов, в том числе в контексте деятельности СМИ, считается безусловным преимуществом, однако в будущем может привести к снижению когнитивных процессов и уровня критического мышления, а также потери множества компетенций, которыми владеют специалисты сегодня.

Несмотря на явные преимущества, важно понимать, что ИИ кардинально меняет жизнь людей, делая одни действия проще и доступнее, а другие труднее и непонятнее. Такой скачок в развитии нейросетей (в данной статье ИИ и нейросети рассматриваются как синонимичные понятия – *прим. авт.*) не позволяет человечеству осознать и обдумать происходящее, заставляя осваивать и усваивать все буквально на ходу. Все это делает современного пользователя уязвимым с точки зрения медиабезопасности, ведь человеческое сознание не успевает отгородиться от всех нововведений, которыми активно пользуются манипуляторы.

ИИ способен создавать настолько качественные фейки, что они неотличимы от оригинала. На это указывает и исследователь А. К. П. Калиан, отмечая, что рост числа убедительных подделок представляет серьезную угрозу для политического устройства страны и конфиденциальности каждого человека [13, p. 1]. Но есть и более серьезные последствия: глобальное распространение дипфейков создает большие риски для стабильности международного порядка [11, с. 101]. И если еще в прошлом году метод анализа теней, отражения в глазах или артефакты генерации действительно позволяли

распознать подделку, то в 2025 г. многие из них уже не актуальны. Конечно, появляются новые способы: частота моргания глаз и пр., но, скорее всего, уже в следующем году и они будут неэффективны.

В научной литературе описаны характерные особенности фейкового контента в СМИ. Так, Е. И. Галяшина пишет о понятии и сущности фейка и фейкинга, отмечает причины распространения фейков [2]. С. Н. Ильченко предлагает типологию фейк-контента и предлагает маркеры распознавания фейков [4]. Способы распознавания лжи в медиатекстах также представлены в труде А. М. Шестериной и И. А. Стернина[8]. Авторы монографии «Фейки: коммуникация, смыслы, ответственность» пишут о коммуникативной природе фейка, семантике фейка и формировании фейкового контента в гипертексте цифровых медиа [12]. Некоторые исследователи предлагают рассматривать искусственный интеллект не только как эффективный инструмент, но и как серьезное «информационное оружие» [9]. Историю появления и развития фейков в медиапространстве предлагает в своей статье Н. В. Манвелов [7]. В зарубежной научной среде также представлено достаточное количество трудов, посвященных проблемам фейков и дипфейков в мировом инфополе. Среди них – материал Дж. Стрея, который говорит о высоком риске столкновения с фейками в журналистских расследованиях, о галлюцинировании нейросетей и опасностях, которые могут из-за этого последовать [15]. Дж. Тандок, Р. Томас и Л. Бишоп анализируют фейковые истории на предмет их полезности и привлекательности для аудитории [16]. М. Волдроп рассматривает фейковые новости как одну из главнейших проблем современного медиапространства [17]. При этом большинство исследователей выказывают серьезные опасения касаемо постоянного совершенствования дипфейков и не видят адекватных позитивных примеров их использования в СМИ.

Логично предположить, что в будущем технологии дипфейка из-за создания ложного видеоролика или фотоизображения могут не только испортить жизнь одному человеку, но привести к массовым беспорядкам, митингам и даже военным столкновениям. Согласно исследованию о фейках АНО «Диалог Регионы», в 2024 г. уникальных дипфейков стало в 5 раз больше, чем в 2023 г. Среди самых распространенных тем – дипфейки глав регионов (35%), публикации, связанные с акцией «Солдат ребенка не обидит» (15%), дипфейки представителей Минобороны РФ (14%). (Исследование по распространению фейковой информации. Ежегодный доклад АНО «Диалог Регионы» // Диалог о фейках 2.0., https://fakes2024.dialog.info/static/files/Исследование_2024.pdf).

Само понятие «дипфейк» (deepfake) сложилось из терминов deep learning (глубокое обучение) и fake (подделка) [6, с. 41]. Данный феномен приписывают появлению новейших цифровых технологий, однако само явление «подмены визуальной реальности» восходит к попыткам подделать фотографии методами двойного экспонирования и ретуши. Так на изображении могло появится то, чего не было в оригинале. Затем появился кинематограф и были придуманы комбинированные съемки и макеты (вспомнить хотя бы нарисованные световые мечи джедаев и картонные звездолеты). Теперь благодаря компьютерам придуманное становится реальным (Иевлев П. Несобственной персоной // Цифровой океан. 2023. № 6 (20)).

Стоит также вспомнить, что даже примитивные возможности фотомонтажа позволяли представлять аудитории изображения пришельцев, НЛО или лохнесского чудовища, а также шантажировать политиков и знаменитостей с целью создания провокаций, сенсаций, давления на общественное мнение. С тех пор слово «дипфейк» имеет

негативную коннотацию, хотя может использоваться и с благими намерениями. Таким образом, *дипфейк представляет собой синтетический контент, в котором фото- или видеоизображение и/или голос человека заменяется на другого.* В итоге получается, что человек может оказаться где-то, где его никогда не было, говорить то, что он на самом деле никогда не говорил, вести себя так, как в реальности он бы никогда себя не повел и т. п.

Дипфейки создаются путем обучения генеративно-состязательной нейронной сети (GAN) [1, с. 165]. Одна нейросеть (генератор) создает изображения, а другая (дискриминатор) – оценивает их. Генератор постоянно создает новые варианты изображений до тех пор, пока дискриминатор не перестанет отличать их от реального изображения. В этом случае люди также не могут увидеть подделку.

Такой подход открывает неограниченные возможности для манипуляторов, поскольку кардинально отличается от подделок прошлых поколений. Раньше для создания убедительной фотографии требовался по-настоящему профессиональный специалист, который тратил много времени и сил на подделку продукта. При этом результат все равно не мог обмануть эксперта. Сейчас генеративный сети доступны каждому, что порождает бесконечное количество дипфейков в медиасфере. В то же время сама медиасфера все больше влияет на человечество. По данным издания The Wall Street Journal количество случаев мошенничества с помощью дипфейков выросло на 700% за последний год (Емельянцева М. Мошенники стали чаще использовать дипфейки для обмана россиян: названы их самые популярные схемы // Men Today. 2024. 10 апр., https://www.mentoday.ru/life/news/10-04-2024/moshenniki-stali-chashche-ispolzovat-dipfeiki-dlya-obmana-rossiyan-nazvany-ih-samye-populyarnye-shemy/). А в России в 2024 г. мошенники стали в семь раз чаще использовать дипфейки в финансовой сфере (Мошенники стали в семь раз чаще использовать дипфейки в секторе финансовых технологий // Искусственный интеллект Российской Федерации, https://ai.gov.ru/mediacenter/moshenniki-stali-v-sem-raz-chashche-ispolzovat-dipfeyki-v-sektore-finansovykh-tekhnologiy/). Как справедливо пишет М. А. Савушкина, «то, что дипфейк из развлекательной технологии превратился в опасное цифровое оружие, – результат деятельности самого человека» [10, с. 54].

Принцип создания дипфейка уже не является ни для кого секретом. Искусственный интеллект объединяет большое количество фотоизображений и делает из них видеозапись. Программа может с высокой точностью определить, как человек будет реагировать и себя вести в определенной ситуации. Иными словами, суть технологии заключается в том, что часть алгоритма детально изучает изображение объекта и пытается его воссоздать, пока другая часть не перестает различать реальное изображение и созданное нейросетью. Дипфейки практически невозможно распознать, поскольку видео как правило отличается высокой степенью реалистичности [5, с. 75].

М. Б. Добробаба дает довольно исчерпывающее определение этому понятию, отмечая, что дипфейки представляют собой технологии изготовления поддельных фото- и видеоизображений, а также аудиозаписей, в основе которых лежит методика компьютерного синтеза. Другими словами, нейросеть переносит черты человека на чужое фото или видео с высокой степенью правдоподобия [3, с. 112], а также генерирует все голосовые записи человека и создает его монолог, который он никогда не произносил. И если в бесплатных версиях ChatGPT или Kandinsky генерацию изображений легко распознать, то в платной версии MidJorney определить фейк практически невозможно.

Нередкими становятся случаи, когда ИИ используется для модуляции голоса и создания поддельного номера телефона, что позволяет обмануть доверчивых граждан и выманить у них большие суммы денег. Подобные примеры демонстрируют серьезные угрозы для неподготовленной аудитории со стороны ИИ. Этим «оружием» могут умело пользоваться манипуляторы и мошенники, преследующие свои цели.

Создание качественных дипфейков требует серьезных компьютерных мощностей и специальных знаний. Они могут использоваться как для создания шуточного видео или рекламного ролика, так и для манипуляции общественным мнением (в том числе через СМИ)[14, p. 69]. Широкое распространение эта технология получила в 2017 г. в США, благодаря разработке технологии «глубокого обучения» (deep learning), в рамках которой ИИ на основе обработки больших данных учится воспроизводить определенные паттерны (модели). На современном этапе дипфейки используются во многих сферах жизнедеятельности. Например, в 2023 г. во Флориде фейковый Сальвадор Дали открыл свою выставку (Museum creates deepfake Salvador Dalí to greet visitors // YouTube, https://www.youtube.com/watch?v=64UN-cUmQMs). Общеизвестны случаи, когда с помощью нейросети рисуют картины, прописывают диалоги в сценариях к сериалам, создают реалистичные фотографии и даже научные тексты. В журналистике дипфейк используют, чтобы скрыть лицо источника, пожелавшего остаться анонимным, и при этом не делать изображение размытым. Однако есть и негативные примеры.

Дипфейк стали активно использовать для создания фейк-ньюс и поддельных видео. Одно из них – скандально известный видеоролик 2018 г., на котором бывший президент США Барак Обама прямо оскорбляет действующего на тот момент президента Дональда Трампа. Этим видео автор ролика, Джордан Пил, продемонстрировал реальную информационную угрозу, которая способна серьезно повлиять на общественное мнение и настроения масс. И теперь публичный деятель, чиновник, политик или представитель шоу-бизнеса может обвинить нейросеть и заявить, что его высказывания – результат работы ИИ, и он там никогда не был и никогда такого не говорил. Насколько возможно это доказать – вопрос сложный и пока только начинающий подниматься в правовом, этическом и научном поле. Однако получить поддержку аудитории таким психологическим приемом уже очевидно можно. Например, продюсер Иосиф Пригожин назвал фейком скандальную аудиозапись своего разговора с бизнесменом Фархадом Ахмедовым (Баласян Л. Иосиф Пригожин отрицает подлинность аудиозаписи беседы с миллиардером Ахмедовым с критикой власти // КоммерсантЪ. 2023. 26 марта, https://www.kommersant.ru/doc/5899921). Однако доказательств этого никто обнаружить не смог. Интересен кейс небольшого американского издания Cody Enterprise, в котором журналист А. Пелчар опубликовал целых семь статей с полностью сгенерированной ИИ прямой речью ньюсмейкеров (Розанова А. Американского журналиста поймали на использовании ИИ и уволили // РБК. 2024. 14 авг., https://www.rbc.ru/life/news/66bcb55f9a7947a6bd7826df). На протяжении двух месяцев репортер придумывал цитаты, пока его не разоблачил коллега из газеты-конкурента.

Представленные выше примеры в полной мере подтверждают существование реальной угрозы для аудитории. При этом в качестве серьезных рисковс позиции использования ИИ в медиа является создание дипфейков преднамеренно или по неосторожности. Социальные сети и блогерский контент является источником информации не только для молодежной аудитории, но и для многих журналистов, которые в погоне за трафиком и популярностью стремятся не столько проверить новость, сколько стать первым и опубликовать «эксклюзив». Сегодня дипфейки могут обмануть даже опытных репортеров.

Другой вариант – когда сама нейросеть ошибается, неверно прочитав обозначения (например, вместо 1–7% указывая 17% или вместо 1925 г. говоря о 2025 г.). Такие фактические ошибки в рамках новостей о котировках акций или финансовых сделках могут привести к глобальным последствиям. В текстах, созданных нейросетью, может отсутствовать контекст происходящего, что также приведет читателя к неверному выводу. Только 52% респондентов исследования, проведенного в 2020 г., смогли отличить контент, сгенерированный нейросетью GPT-3, от текста, созданного человеком (Искусственный интеллект в цифрах и фактах // РБК. 2024. № 01-02 (178)).

В октябре 2023 г. по интернету распространился видеоролик, на котором известная шведская активистка Грета Тунберг рекламировала свою новую книгу «Веганские войны» на BBC (телеканал заблокирован на территории Российской Федерации). Она якобы заявила, что необходимо перейти на экологические танки и вооружение, а вместо нынешних ручных гранат использовать веганские, чтобы ни одно животное не пострадало (O'Rourke C. Greta Thunberg urged people to use "vegan grenades" because "no animals should have to give their life for all this mayhem and chaos" // PolitiFact. 2023. 25 Oct.,https://www.politifact.com/factchecks/2023/oct/25/viral-image/no-greta-thunberg-didnt-urge-people-to-use-vegan-h/). СМИ сразу распознали фейк и указали на поддельное видео, которое было изменено. В частности, журналисты отметили, что речь активистки не соответствовала движениям ее рта, текст ее выступления был изменен. Еще один интересный пример связан с экспериментом, в ходе которого 436 зрителей смотрели дипфейки, в том числе фрагменты фильма «Сияние» с Брэдом Питом и «Капитан Марвел» с Шарлиз Терон (Оптический обман // Цифровой океан. 2023. № 5 (19)). Целью было изучение процесса ложных воспоминаний. По итогу 49% (почти половина) поверили в достоверность показанных им фрагментов.

В ходе исследования было проанализировано 69 дипфейков, появившихся в российском информационном пространстве. Материал был собран методом сплошной выборки и содержал в себе публикации, которые так или иначе были упомянуты в СМИ (на телеканалах «Первый канал», НТВ и «Россия 1», платформе «Смотрим», интернет-изданиях «Лента.ру» и «Газета.ру», сайтах информационных агентств «РИА Новости» и ТАСС), т. е. имели общественный резонанс. Хронологические рамки исследования: 2020–2025 гг.

На основе представленной выше информации и анализа собранных материалов, можно типологизировать дипфейки по разным категориям. *По формату подачи выделим:*

¾ фотодипфейки

Такие фотографии являются результатом генерации нескольких изображений, которые были загружены в память нейросети. Чем больше будет загружено изображений, например, конкретного человека, тем выше вероятность создания его фотодипфейка. Это самый старый тип подделок, который уже вряд ли может как-то убедить аудиторию. То, что фотографию можно подделать, сейчас известно любому человеку, который является пользователем интернета.

¾ видеодипфейки

Видеодипфейки распознать гораздо сложнее, чем подобный контент в формате фотографий. Такие ролики выглядят максимально реалистично (особенно, если есть синхрон речи с движением губ и имеются блики в глазах), да и аудитория на психологическом уровне гораздо больше доверяет видеоизображению, по привычке полагая, что видео является прямым и неоспоримым доказательством. Во многом из-за

этого россияне имеют высокий риск стать жертвой обмана при видеообращениях от родственников или своих руководителей.

¾ аудиодипфейки

Аудиодипфейки могут стать серьезным оружием мошенников в рамках обмана граждан и выманивания у них денежных средств. Кроме того, сейчас эта технология активно используется для озвучивания пиратских копий аудиокниг, когда диктор не получает гонорар, а его голос при этом используется для озвучивания литературного произведения. Кроме того, все чаще появляется информация о случаях, когда подделывают какие-то комментарии медийных личностей (позже оказывается, что такого комментария никто не давал). Таким образом, становится все сложнее разобраться в потоке приходящей информации и не стать жертвой обмана.

По целевым установкам дипфейки можно разделить на:

¾ *политические*

Цель – дискредитация оппонента путем разрушения его репутации с помощью подделок, которые либо демонстрируют в невыгодном свете политику того или иного деятеля, либо самого этого деятеля выставляют в неприглядном свете.

Самым известным и нашумевшим примером являются фотографии и видеоролики с арестом Дональда Трампа и его выступлениями в роли заключенного в оранжевой форме. Материалы были сделаны настолько реалистично, что многие поверили и способствовали распространению этих материалов.

Рис. 1. Дипфейк Дональда Трампа (Нейросеть создала фотохронику задержания Дональда Трампа // Смотрим. 2023. 21 марта, https://smotrim.ru/article/3261001)

В рамках предвыборной кампании 2024 г. в США команда Д. Трампа также активно использовала нейросети для дискредитации политики Д. Байдена. На YouTube активно распространялся ролик Beat Biden, полностью сгенерированный ИИ. В нем аудитория должна была увидеть, какие проблемы ждут США при переизбрании действующего президента на второй срок.

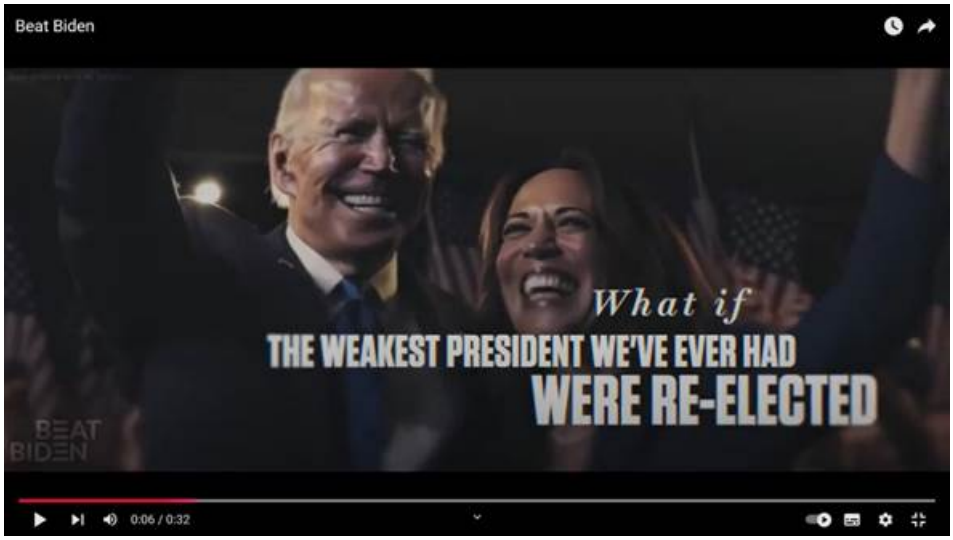

Рис. 2. Видеоролик, созданный нейросетью (Beat Biden // YouTube. 2024, https://www.youtube.com/watch?v=kLMMxgtxQ1Y&t=6s)

В России дипфейки с политиками появляются все чаще. Опасность в том, что пользователи могут принять их за агитационные материалы или политическую рекламу, даже не подумав проверить на предмет подлога или подделки. В августе 2021 г. в интернете появился дипфейк, который за считанные часы стал вирусным. В видеоролике С. Лавров (министр иностранных дел РФ), Д. Проценко (главврач больницы № 40 в Коммунарке) и С. Шойгу (бывший на тот момент министром обороны РФ) советуются с обычным россиянином на темы ядерной войны, вакцинации и прочих важнейших государственных вопросов (В Сети появился вирусный ролик с дипфейками Лаврова, Проценко и Шойгу. Технологию все чаще используют в рекламе и политической агитации // Moscow Daily News. 2021. 27 авг., https://www.mn.ru/smart/v-seti-poyavilsya-virusnyj-rolik-s-dipfejkami-lavrova-proczenko-i-shojgu-dipfejki-vse-chashhe-ispolzuyut-v-reklame-i-politicheskoj-agitaczii). В конце видео главный герой (рядовой гражданин нашей страны) так напуган, что быстрее спешит на выборы, чтобы проголосовать и переложить принятие важных решений в надежные руки. Сразу после выхода ролика партия «Единая Россия» сообщила о своей непричастности к его созданию, что сразу позволило отнести данный контент к категории дипфейка.

Таким образом, политический дипфейк может создаваться не только для дискредитации конкурентов (как это показано на примере американских кейсов), но и для агитации и поддержки определенного кандидата, партии или организации. Все зависит от целей создателя.

¾ *развлекательные*

Такие дипфейки носят юмористический характер, а их создатели не преследуют цель обмануть доверчивых пользователей или повлиять на их мнение о чем-то. Однако здесь важно помнить о том, что даже самый безобидный на первый взгляд дипфейк может нести в себе потенциальную опасность ровно до тех пор, пока его авторы четко не обозначат, что это подделка и предлагаемый контент не соответствует действительности.

Целью подобных дипфейков может стать желание порадовать свою аудиторию, попробовать сделать что-то необычное, стремление выделиться или чем-то запомниться. Как правило, успешные развлекательные дипфейки носят вирусный характер и могут на протяжении длительного времени передаваться по каналам социальных медиа. Такой контент можно встретить в социальных сетях или на сайтах, посвященных

юмористической тематике. Один из популярных вариантов таких дипфейков – создание роликов, где лицо актеров в известных фильмах заменяют лицом других не менее знаменитых актеров. Например, в известном фильме Э. Рязанова «Ирония судьбы, или с легким паром!» лицо Барбары Брыльска заменили на лицо голливудской актрисы Марго Робби, а Андрея Мягкова – на Дэниэла Крейга (Ирония с Голливудскими актерами // Pikabu, https://pikabu.ru/tag/Deepfake%2CЮмор). Данный кейс вряд ли можно назвать вредоносным, поскольку в описании ролика сразу отмечено, что это технология дипфейка, да и само содержание ролика имеет исключительно развлекательный посыл.

¾ *коммерческие*

Цель таких дипфейков – привлечь внимание и получить прибыль. Как правило, такой контент носит рекламный характер. Одним из первых примеров стала реклама продуктов Сбера с Жоржем Милославским из фильма «Иван Васильевич меняет профессию». В ролике 2020 г. герой попадает в современную реальность и с помощью возможностей сервисов Сбера быстро получает все, что ему нужно.

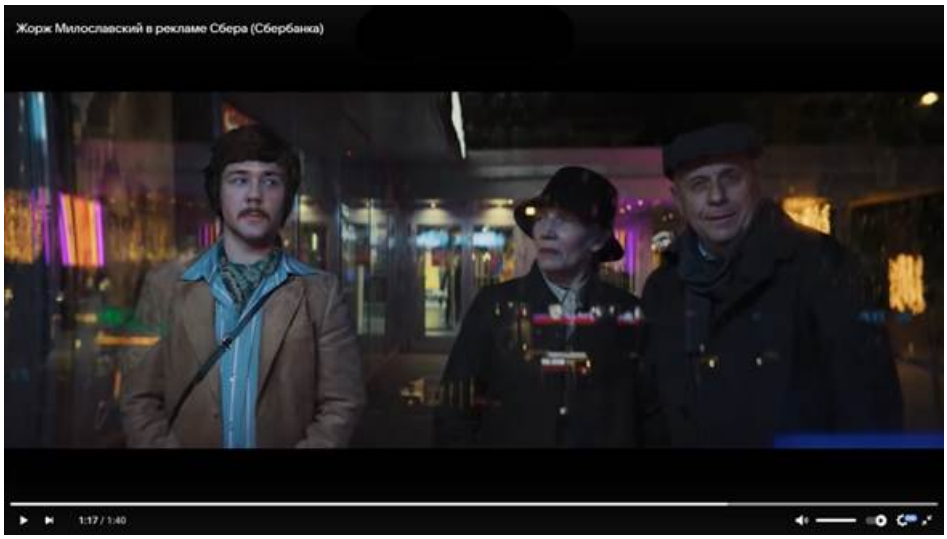

Рис. 3. Реклама Сбера с цифровым аватаром (Жорж Милославский в рекламе Сбера (Сбербанка) // VK Видео. 2021, https://vkvideo.ru/video116656652_456240165?ref_domain=yastatic.net)

Подобный материал способен значительно повысить узнаваемость бренда и обратить на себя внимание разнообразной аудитории. Такие ролики могут использовать изображение или голос знаменитости, а также быть созданы по сценарию, созданному ИИ.

¾ *художественные*

Возможности дипфейков в последние несколько лет стали активно использовать кинематографисты, которые могут состарить или омолодить актера, придать дублерам максимальное сходство с главным героем, синхронизировать движения губ при дубляже перевода. Более того, в рамках создания художественных или документальных картин сейчас «оживляют» давно умерших звезд и видных деятелей культуры и искусства прошлых лет и даже столетий.

В 2023 г. режиссер и продюсер А. Жигалкин анонсировал съемку фильма «Володя» о Владимире Высоцком. Главную роль, по задумке создателя, должны будут одновременно сыграть актер А. Шпагин и искусственный интеллект, который смоделирует на экране лицо певца. В том же году появилось сообщение, что ИИ воссоздаст голос Эдит Пиаф для кинокартины о жизни знаменитости. А в декабре 2023 г. ТАСС сообщил, что

нейросети синтезируют известный на весь мир голос Юрия Левитана к 110-летию его рождения. Для обучения ИИ был использован архив записей голоса диктора из Госфильмофонда. В 2024 г. в СМИ опубликовали сообщение, что «Первый канал» планирует снять проект о Штирлице по повестям Юлиана Семенова. Образ Вячеслава Тихонова будет воссоздан с помощью технологий ИИ (Мамиконян С. Первый канал воссоздаст с помощью ИИ образ Вячеслава Тихонова в роли Штирлица // Forbes. 2024. 05 нояб., https://www.forbes.ru/forbeslife/524484-pervyj-kanal-vossozdast-s-pomos-u-ii-obraz-vaceslava-tihonova-v-roli-stirlica). Иными словами, тренд на создание цифровых двойников известных актеров прошлого продолжает сохраняться.

Насколько такие дипфейки будут пользоваться спросом пока непонятно. Это может привлечь аудиторию эффектом новизны, но вряд ли надолго, ведь заменить реального актера (мимику, реакцию, выражение глаз и пр.) на данном этапе развития ИИ не в состоянии. Кроме того, опросы показывают, что россияне относятся к ИИ-контенту крайне скептически. А потом здесь есть риск и для живых актеров, которые теряют возможности проявить себя из-за цифровых образов давно умерших знаменитостей. Следовательно, в 2025 г. перспектива использования таких дипфейков выглядит сомнительной.

¾ *манипуляционные*

Самый опасный тип дипфейков, который подразумевает целенаправленный обман пользователя ради достижения своей цели. Как правило, такой контент является вредоносным, деструктивным и часто приводит к неприятным последствиям. Именно дипфейки стали причиной многочисленных обманов, из-за которых россияне теряют свои деньги, имущество и совершают противоправные действия. Контент в социальных медиа становится по-настоящему опасным, поскольку в 2025 г. уже не единичны примеры, когда мошенники подделывают голоса родственников, создают дипфейк в формате видео с родными людьми, совершают звонки с номеров родственников, тем самым очень убедительно манипулируя сознанием человека и не давая даже осознать происходящее. Как правило, злоумышленники в очень оперативном режиме вынуждают переводить на незнакомые счета все свои сбережения. Единственным способом защиты в этом случае является завершение разговора и самостоятельный звонок своему родственнику или другу.

¾ *познавательные или просветительские*

В целях помощи широкой аудитории разобраться, каким образом создаются дипфейки и в чем суть этой технологии, IT-специалисты и хакеры выпускают разъясняющие видеоролики, где подробно и доступно описан сам процесс создания такого материала. Например, довольно известным является видеролик, где сам автор наглядно демонстрирует, как создает дипфейк с Морганом Фрименом (см. Рис. 4).

Рис. 4. Ролик с демонстрацией возможностей ИИ(Не все, что мы видим – реально. Дипфейк с лицом Моргана Фримена. // Одноклассники. 2022, https://ok.ru/video/4717712050766)

В ролике мужчина в нижней части рекламы произносит речь, которую полностью копирует Морган Фримен (вверху экрана), при этом также идентичны мимические реакции и эмоциональные выражения. Таким образом, зритель собственными глазами видит, насколько такая технология может создавать реалистичные изображения человека.

В целом, дипфейки представляют собой актуальную информационную угрозу, поскольку в отличие от фотографий, коррекция которых не является чем-то уникальным, видеозаписям пользователи склонны по-прежнему доверять. Поэтому дипфейки в видеоформате с большой долей вероятности будут восприняты как подлинные, а благодаря социальным сетям могут быть растиражированы за считанные минуты.

В 2025 г. на первый план в связи с указанными угрозами выступила безопасность детей в интернете. Эта часть аудитории является наиболее уязвимой к приемам манипуляции и психологического воздействия. В январе 2025 г. известный сервис TikTok стал блокировать так называемые фильтры красоты для аудитории младше 18 лет (Розанова А. TikTok запретит подросткам пользоваться бьюти-фильтрами. Зачем это нужно // РБК Life. 2024. 27 нояб., https://www.rbc.ru/life/news/6746fe379a794770e0b615e8). Подростки больше не смогут использовать маски, делающие лицо буквально совершенным. Проблема безопасности детей в интернете, особенно их психологическое состояние, которое является в юном возрасте неустойчивым, – предмет для беспокойства во всех странах мира. В 2023 г. в Великобритании появилась информация, что власти собираются запретить британцам младше 16 лет вообще использовать социальные сети (В Британии могут запретить детям до 16 лет пользоваться соцсетями // Известия. 2023. 15 дек., https://iz.ru/1621117/2023-12-15/v-britanii-mogut-zapretit-detiam-do-16-let-polzovatsia-sotcsetiami). В России 6 декабря 2023 г. Госдума приняла законопроект об ограничении использования телефонов и других цифровых гаджетов в школе (Федеральный закон от 19 декабря 2023 г. № 618-ФЗ «О внесении изменений в Федеральный закон "Об образовании в Российской Федерации"» // Гарант.ру, https://www.garant.ru/products/ipo/prime/doc/408131681/), что, по мнению законодателей, должно помочь воспитать вдумчивых людей, способных к критическому мышлению. Однако, в ситуации масштабных обманов с помощью реалистичных подделок изображений и голоса, такие ограничения имеют довольно ограниченную меру.

Вероятно, что в ближайшем будущем закон о дипфейках будет создан и поможет защитить и взрослое население, и подрастающее поколение. Но пока в 2025 г. главным оружием борьбы и самозащиты выступает только критическое мышление. Чем в большей опасности себя чувствуют люди, тем более они подвержены влиянию. Такие пользователи менее критичны, меньше сил тратят на проверку информации. Даже рациональные потребители медиаконтента могут быть введены в заблуждение, чем умело пользуются манипуляторы. В этом случае определенный алгоритм позволит убедиться в подлинности получаемых данных:

1) прочесть полностью материал (заголовок не всегда отражает содержание или суть текста);

2) определить автора и дату публикации (это крупное СМИ или чей-то авторский блог, насколько автор текста авторитетен и знает тему);

3) проверить адресную строку (это могут быть «клоны», которые копируют страницы настоящих медиа, изменяя букву или добавляя какой-то знак пунктуации, или фейковые аккаунты, которые также отличаются буквально одним символом в названии);

4) найти и изучить первоисточник информации;

5) в фейковых материалах большое количество эмоциональных высказываний, субъективных суждений, оценочных выражений (упор на сенсацию или эксклюзив). Как правило, в журналистских материалах на первое место выходит фактическая информация, а эмоциональные высказывания играют второстепенную роль.

Однако если речь идет о дипфейках, этих правил может быть недостаточно. При сгенерированном фото или видео стоит обращать внимание на качество изображения, естественность телодвижений, блики в глазах или на очках, естественность освещения и фоновых цветов и пр. Иными словами, увидеть искусственно созданные кадры может только крайняя внимательность и целенаправленный поиск.

В частности, *в качестве опознавательных признаков дипфейка (которые могут стать первостепенными при определенных обстоятельствах) важно выделить:*

¾ несоответствие голоса и движений губ с произносимыми словами (если слова и движения не совпадают, это может стать признаком подделки), а также монотонность произнесения речи, отсутствие логических пауз, ошибки в ударениях, использование нехарактерных для носителей языка словесных конструкций;

¾ артефакты и неестественное освещение (если есть «размытые» края объектов или свет не выглядит естественным, это может быть признаком цифровой обработки изображения);

¾ неестественные движения (резкие движения или, наоборот, замедленные реакции могут стать доказательством подделки);

¾ несоответствия и искажения (отсутствие теней, разрывы линий, явно лишние элементы, игнорирование законов физики, повторяющиеся конструкции или элементы на изображении и т. п.). Здесь имеет смысл обратить особое внимание на аксессуары и одежду как самые «слабые» стороны работы ИИ – сережка может быть только в одном ухе, непропорциональный воротник рубашки, странные пуговицы, несоответствие цветов и пр.;

¾ нереальные лица (разные глаза, количество пальцев на руках, асимметрия в деталях, нереалистичные зубы, сюрреалистический фон, самая частая ошибка нейросетей – это уши: зачастую второе ухо может вовсе исчезнуть или быть неподходящим по размеру);

¾ отсутствие физического взаимодействия (если на кадрах два объекта, которые должны реагировать друг на друга, но этого не происходит, это может стать признаком генерации видеозаписи).

Так мы возвращаемся к важнейшему правилу – проверке источника. Подобная процедура поможет узнать распространителя сомнительного контента, определить, насколько этот распространитель может считаться надежным. Для этого стоит изучить метаданные файла – узнать, где, когда и кем был создан материал.

На данный момент существуют и специальные программы, позволяющие распознать генерацию (интересный феномен: распознать ИИ способен только сам ИИ). Например, ресурс Optic AI or Not определяет, сгенерировано или нет изображение (Optic AI or Not, https://www.aiornot.com/). Сервис является бесплатным и интуитивно понятным. Все, что требуется от пользователя, – загрузить изображение на портал. При этом программа получила мало положительных оценок, поскольку при тестировании журналистами делала ошибки и определяла сгенерированные фотографии как реальные и наоборот. Подобные прецеденты позволяют сделать вывод, что на современном этапе подобным проектам нельзя полностью доверять, слишком велик риск ошибок.

Специально созданный для выявления подделок нейросетевой алгоритм LFCC-LCNN смог распознать 100% аудиофейков, загруженных в него (Иевлев П. Люди недостаточно хорошо различают голосовые фейки // Цифровой океан. 2023. 05 авг., https://digitalocean.ru/n/lyudi-nedostatochno-horosho-razlichayut-golosovye-fejki).
Согласно проведенному британскими исследователями эксперименту, специально обученные люди смогли распознать фейки лишь чуть более чем в 70% случаев, в то время как алгоритм не сделал ни одной ошибки.

OpenAI запустила инструмент для выявления изображений, сгенерированных ее нейросетью DALL-E3. По словам самих разработчиков, система способна выявить дипфейк с точность 98%, однако процент будет гораздо ниже, если изображение было изменено (обрезано, подвергалось изменению цвета или был загружен скриншот) (Seetharaman D. OpenAI Says It Can Now Detect Images Spawned by Its Software – Most of the Time // The Wall Street Journal. 2024. 07 May, https://www.wsj.com/tech/ai/openai-says-it-can-now-detect-images-spawned-by-its-softwaremost-of-the-time-83011149).
Данный проект появился как ответ на резкий рост различных дипфейков из-за активизации темы предвыборных кампаний 2024 г. в разных странах.

В 2022 г. компания Intel анонсировала свой продукт под названием FakeCatcher, который выявляет дипфейк на основе анализа цветовых пульсаций подкожных вен лица (Intel FakeCatcher Технология распознавания дипфейков // Tadviser, https://www.tadviser.ru/index.php/Продукт:Intel_FakeCatcher_Технология_распознавания_дипфейков). Детектор способен фиксировать незаметные глазу изменения тона, переводить их в цветовую карту и определять, настоящий ли это человек. Однако официального внедрения данной программы на рынок пока не начиналось. А компания Sensity разработала онлайн-платформу для автоматической идентификации дипфейков. В основе работы программы – анализ кадров изображения на основе собственной базы, которая содержит несколько миллионов изображений, определенных как сгенерированные. Система обучена выявлять признаки генерации, которые используют

нейросети.

В НИТУ МИСИС отечественные ученые разработали нейросеть, которая представляет собой поисковик фейковых лиц на фото или видео (Бунина В. Создана нейросеть для идентификации дипфейков на фото и видео // Газета.ru. 2023. 16 окт., https://www.gazeta.ru/tech/news/2023/10/16/21512545.shtml). При разработке было использовано 16,5 тысяч изображений как подлинных, так и фейковых, которые стали базой для обучения нейросети.

С 2020 г. на базе Университета Лобачевского функционирует сетевой образовательный проект «#Студфактчек», который ориентирован на проверку информации в СМИ и социальных медиа (#Студфактчек, https://studfactcheck.ru). На сайте проекта можно прочитать подробный разбор материалов, которые члены команды фактчекеров проверяют на предмет достоверности. Это важный и интересный проект, способный сориентировать пользователя прежде всего в методологических способах проверки фактов. «#Студфактчек» по принципу работы напоминает известный портал «Лапша.медиа», где сотрудники также выявляют фейки и дипфейки и предупреждают о них интернет-пользователей.

Описанные выше примеры демонстрируют высокий запрос на создание программ, способных выявлять дипфейки и бороться с ними. Вполне возможно, что в скором будущем такие программы станут обязательными для установки на всех гаджетах, появятся специальные мобильные приложения, использование которых будет также естественно, как приложение Telegram или Wildberries. И в порядке вещей будет автоматическая проверка любого фото- и видеоконтента на предмет достоверности. Есть также вариант, что такие «распознаватели» станут обязательной частью видеоплатформ и фотобанков, и при обнаружении генерации контент будет автоматически промаркирован.

Но это только возможные перспективы. Конечно, какие-то крупные и резонансные дипфейки будут разоблачены публично, однако это лишь незначительная часть той неправды, с которой сталкивается ежедневно человек. Именно критическое осмысление получаемых данных и постоянная бдительность при потреблении медиаконтента – то, что остается самым эффективным способом самозащиты в условиях непрерывно увеличивающегося количества информационных угроз.

Библиография

1. Батоев В. Б., Пучнин А. В. Использование технологии deepfake в преступной деятельности: проблемы противодействия и пути их решения // Вестник ВИ МВД России. 2023. № 1. С. 165-169. EDN: MIUJNO.
2. Галяшина Е. И. и др. Фейковизация как средство информационной войны в интернет-медиа: научно-практическое пособие. М.: Блок-Принт, 2023. EDN: NTUNGW.
3. Добробаба М. Б. Дипфейки как угроза правам человека // Lex Russica. 2022. № 11 (192). С. 112-119. DOI: 10.17803/1729-5920.2022.192.11.112-119. EDN: XMHEAJ.
4. Ильченко С. Н. Фейк-контроль, или Новости, которым не надо верить: как нас дурачат СМИ. Ростов н/Д: Феникс, 2021.
5. Колесникова Е. В. Технология deepfake в журналистике // Мир современных медиа: новые возможности и перспективы: сб. научных трудов / под общ. ред. Д. В. Неренц. М.: Знание-М, 2022. С. 74-77. EDN: CCYEHV.
6. Лукина Ю. В. Использование дипфейков в общественно-политической жизни // Русская политология. 2023. № 2 (27). С. 41-48. EDN: CEXDQA.
7. Манвелов Н. В. Понятие "фейк" в медиакоммуникациях – история и современные

подходы к проблеме // Коммуникация – дискурс – дискурсивные практики: Сборник научных трудов. М.: РАНХиГС, 2023. С. 148-159.
8. Маркеры фейка в медиатекстах: учебно-методическое пособие / И. А. Стернин, А. М. Шестерина, К. И. Грибанова [и др.]. Воронеж: РИТМ, 2020.
9. Неренц Д. В. Специфика применения искусственного интеллекта в современном медиапространстве // Litera. 2024. № 8. С. 186-198.
10. Савушкина М. А. Дипфейк как цифровое оружие гибридной войны // Вестник ОмГУ. 2024. № 4. С. 45-54.
11. Фалалеев М. А., Ситдикова Н. А., Нечай Е. Е. Дипфейк как феномен политической коммуникации // Вестник ЗабГУ. 2021. № 6. С. 101-106. DOI: 10.21209/2227-9245-2021-27-6-101-106. EDN: ZBZRKO.
12. Фейки: коммуникация, смыслы, ответственность. Коллективная монография / С. Т. Золян, Н. А. Пробст, Ж. Р. Сладкевич, Г. Л. Тульчинский; под ред. Г. Л. Тульчинского. СПб.: Алетейя, 2021.
13. Kalyan A.K.R. A review on Ethical and Legal Challenges of Deepfake Technology // International Journal Of Scientific Research In Engineering And Management. 2025. No 09 (04). Pp. 1-9.
14. Marconi F. Newsmakers: Artificial Intelligence and the Future of Journalism. NY: Columbia University Press, 2020.
15. Stray J. Making Artificial Intelligence Work for Investigative Journalism // Digital Journalism. 2019. No 7 (8). Pp. 1076-1097.
16. Tandoc Jr. E., Thomas R., Bishop L. What Is (Fake) News? Analyzing News Values (and More) in Fake Stories // Media and Communication. 2021. Vol. 9. No 1. Pp. 112-123.
17. Waldrop M. News Feature: The genuine problem of fake news // PNAS. 2017. No 114 (48). Pp. 12631-12634.

Результаты процедуры рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.
Со списком рецензентов издательства можно ознакомиться здесь.

Вариант статьи представленный к публикации касается достаточно сложной, но актуальной темы, автор ориентирован на анализ проблемы «дипфейков» в информационном пространстве. Причем автор работы обозначает уже в заглавии, что это есть «угроза», которая, так или иначе, разрушает традиционное разрешение и речевых, и не речевых ситуаций. Стоит согласиться, что «разговоры об искусственном интеллекте и его возможностях сегодня не просто не прекращаются, но многократно множатся. Его обсуждают ученые, практикующие специалисты, работники IT-сферы, государственные служащие», «автоматизация рутинных процессов, о которых уже написано и сказано множество слов, в том числе в контексте деятельности СМИ, считается безусловным преимуществом, однако в будущем может привести к снижению когнитивных процессов и уровня критического мышления, а также потери множества компетенций, которыми владеют специалисты сегодня». Таким образом, верификация вопроса представлена весьма удачно, именно такой тон и характерен всему изысканию. Методология анализа, на мой взгляд, вполне актуальна, ибо автор стремится к максимальной объективации вопроса в рамках систематизации уже имеющихся данных. Считаю, что это удается сделать полновесно, целостно; цитатный фон / ссылочная база достаточны. Однако, можно было использовать традиционный вариант отсылки – «…» [1, с. 123]. Не лишена работы и эффекта возможного диалога с оппонентами, сделано это выверено, точно: например, «В научной литературе описаны характерные особенности

фейкового контента в СМИ. Так, Е. И. Галяшина пишет о понятии и сущности фейка и фейкинга, отмечает причины распространения фейков [2]. С. Н. Ильченко предлагает типологию фейк-контента и предлагает маркеры распознавания фейков [4]. Способы распознавания лжи в медиатекстах также представлены в труде А. М. Шестериной и И. А. Стернина [8]» и т.д. При этом представлен как отечественный, так и зарубежный опыт, что вполне значимо для объективации научной новизны данного труда. Стиль работы соотносится с научным типом, термины и понятия, которые используются по ходу статьи унифицированы: например, «само понятие «дипфейк» (deepfake) сложилось из терминов deep learning (глубокое обучение) и fake (подделка). Данный феномен приписывают появлению новейших цифровых технологий, однако само явление «подмены визуальной реальности» восходит к попыткам подделать фотографии методами двойного экспонирования и ретуши. Так на изображении могло появится то, чего не было в оригинале». Привлекает в исследование введение статистических данных, отсылка на открытость информации не вызывает сомнений. Укажу, что аналитическая составляющая в работе весьма умело выдержана, аргументация – есть необходимое условие научного труда: «принцип создания дипфейка уже не является ни для кого секретом. Искусственный интеллект объединяет большое количество фотоизображений и делает из них видеозапись. Программа может с высокой точностью определить, как человек будет реагировать и себя вести в определенной ситуации. Иными словами, суть технологии заключается в том, что часть алгоритма детально изучает изображение объекта и пытается его воссоздать, пока другая часть не перестает различать реальное изображение и созданное нейросетью». Материал оригинален, интересен, думаю, что он будет вполне адекватно воспринят читателями, да и воспринимать его следует как некий импульс для формирования новых изысканий смежно-тематической направленности. Данных, которые были проанализированы, вполне достаточно; автор отмечает, что «в ходе исследования было проанализировано 69 дипфейков, появившихся в российском информационном пространстве. Материал был собран методом сплошной выборки и содержал в себе публикации, которые так или иначе были упомянуты в СМИ (на телеканалах «Первый канал», НТВ и «Россия 1», платформе «Смотрим», интернет-изданиях «Лента.ру» и «Газета.ру», сайтах информационных агентств «РИА Новости» и ТАСС), т. е. имели общественный резонанс. Хронологические рамки исследования: 2020–2025 гг.». Причем источники разные, что дает возможность полновесно оценить «угрозу от дипфейков». Выводы по тексту созвучны основной части, противоречий в этой области нет. Автор в финале тезирует, что «какие-то крупные и резонансные дипфейки будут разоблачены публично, однако это лишь незначительная часть той неправды, с которой сталкивается ежедневно человек. Именно критическое осмысление получаемых данных и постоянная бдительность при потреблении медиаконтента – то, что остается самым эффективным способом самозащиты в условиях непрерывно увеличивающегося количества информационных угроз». Считаю, что тема работы раскрыта, но исследование в указанном русле может быть продолжено. Общие требования издания учтены; фактическая правка текста излишня. Рекомендую статью «Дипфейк как одна из главных информационных угроз XXI века» к открытой публикации в журнале «Филология: научные исследования».