**ПОЯВЛЕНИЕ ТЕХНОЛОГИИ DEEPFAKE.**

(А.С. Ходунов - юрист, судебный эксперт)

*«Ситуация развивается быстрее, чем я думал. Вскоре наступит момент, когда мы больше не сможем обнаруживать дипфейки, поэтому нам придется искать другие типы решений».*

*Хао Ли*

*Пионер Deepfake*

**АБСТРАКТНЫЙ**

Новые цифровые технологии затрудняют различие между реальными и фейковыми СМИ. Одним из последних событий, усугубляющих эту проблему, является появление дипфейков — гиперреалистичных видеороликов, в которых применяется искусственный интеллект (ИИ), чтобы изобразить, как кто-то говорит или делает то, чего никогда в жизни не происходило.

В сочетании с охватом и скоростью социальных сетей убедительные дипфейки могут быстро охватить миллионы людей и оказать негативное влияние на наше общество. Хотя научные исследования по этой теме немногочисленны, в своей статье я попытаюсь проанализировать 84 общедоступные новостные статьи в Интернете, чтобы выяснить, что такое дипфейки и кто их производит, каковы преимущества и угрозы технологии дипфейков, какие существуют примеры дипфейков и как с ними бороться.

Результаты показывают, что, хотя дипфейки представляют собой значительную угрозу нашему обществу, политической системе и бизнесу, с ними можно бороться с помощью законодательства и регулирования, корпоративной политики и добровольных действий, образования и обучения, а также разработки технологий обнаружения дипфейков, контента, аутентификации и предотвращении дипфейков. Моя работа представляет собой обзор дипфейков и предоставляет предпринимателям в области кибербезопасности и искусственного интеллекта бизнес-возможности в борьбе с подделками в СМИ и фейковыми новостями.

**ВВЕДЕНИЕ**

В последние годы фейковые новости стали проблемой, представляющей угрозу общественному дискурсу, человеческому обществу и демократии.

Под фейковыми новостями понимают вымышленный новостной контент, созданный с целью обмана общественности. Ложная информация быстро распространяется через социальные сети, где она может повлиять на миллионы пользователей. В настоящее время каждый пятый пользователь Интернета получает новости через YouTube, уступая только Facebook[[1]](#footnote-1).

Рост популярности видео подчеркивает необходимость в инструментах для подтверждения подлинности медиа- и новостного контента, поскольку новые технологии позволяют убедительно манипулировать видео. Учитывая легкость получения и распространения дезинформации через платформы социальных сетей, становится все труднее понять, чему доверять, что приводит, среди прочего, к пагубным последствиям для принятия обоснованных решений.

Действительно, сегодня мы живем в эпоху, которую некоторые называют «постправдой», которая характеризуется цифровой дезинформацией и информационной войной, возглавляемой злоумышленниками, проводящими кампании ложной информации с целью манипулирования общественным мнением.

Последние технологические достижения позволили легко создавать то, что сейчас называют «дипфейками», гиперреалистичными видеороликами с использованием замены лиц, которые не оставляют следов манипуляций.

**Дипфейки** - это продукт приложений искусственного интеллекта (ИИ), которые объединяют, комбинируют, заменяют и накладывают друг на друга изображения и видеоклипы для создания поддельных видеороликов, которые выглядят аутентичными.

Технология Deepfake может генерировать, например, юмористическое, порнографическое или политическое видео, в котором человек говорит что-либо, без согласия человека, чей образ и голос задействованы.

Фактором, меняющим правила игры в дипфейках, является масштаб и сложность задействованных технологий, поскольку практически любой, у кого есть компьютер, может создавать фейковые видеоролики, практически неотличимые от подлинных медиа.

В то время как ранние примеры дипфейков были сосредоточены на политических лидерах, актрисах, комиках и артистах, чьи лица вплетены в порнографическое видео, в будущем дипфейки, вероятно, будут все больше и больше использоваться для порномести, издевательств и фейковых видео, доказательств в судах, политического саботажа, террористической пропаганды, шантажа, манипулирования рынками и фейковыми новостями.

Хотя распространять ложную информацию легко - исправлять записи и бороться с дипфейками намного сложнее. Чтобы бороться с дипфейками, нам необходимо понимать дипфейки, причины их существования и лежащие в их основе технологии.

Однако, научные исследования лишь недавно начали заниматься цифровой дезинформацией в социальных сетях. Поскольку дипфейки появились в Интернете только в 2017 году, научная литература по этой теме немногочисленна.

Таким образом, целью моего исследования является обсуждение того, что такое дипфейки и кто их производит, каковы преимущества и угрозы технологии дипфейков, некоторые примеры современных дипфейков и способы борьбы с ними.

При этом в данной работе анализируется ряд новостных статей о дипфейках, взятых с сайтов новостных СМИ. Исследование вносит вклад в зарождающуюся литературу о фейковых новостях и дипфейках, предоставляя всесторонний обзор дипфейков, а также внедряя возникающую тему в академические дебаты, которые также определяют варианты борьбы с дипфейками для политиков, журналистов, предпринимателей и других лиц.

Данный материал организован следующим образом: после введения в исследование объясняется сбор данных и анализ новостных статей. Затем в исследовании предлагаются четыре раздела, в которых рассматриваю дипфейки, каковы потенциальные преимущества технологии дипфейков, кто является участниками создания дипфейков, а также угрозы дипфейков нашему обществу, политическим системам и бизнесу.

Далее в двух разделах привожу примеры дипфейков и обсуждаю четыре возможных механизма борьбы с дипфейками. Наконец, исследование завершаю выводами, ограничениями и предложениями для будущих исследований.

**МЕТОД**

Это исследование опирается на появляющуюся научную литературу и общедоступные новостные статьи о дипфейках. Всего мной было собрано 84 статьи с интернет сайтов, 11 новостных компаний с целью проведения эмпирического анализа того, как средства массовой информации обсуждают дипфейки.

Все статьи, посвященные дипфейкам, были написаны на английском языке и опубликованы в 2022–2023 годах. Их нашел через поиск Google News, по ключевым словам, «Deepfake», «Deep fake» и соответствующим формам множественного числа.

После того как статья была найдена, аналогичный поиск был выполнен с использованием собственной функции поиска новостного веб-сайта, чтобы найти больше статей из этого конкретного источника СМИ.

Фокус выбранных средств массовой информации варьировался от общих ежедневных новостей до новостей бизнеса или технологий. Набор данных включает от 2 до 16 новостных статей о дипфейках от каждой новостной компании. Статьям был присвоен короткий идентификатор для цитирования, а затем они были проанализированы с помощью контент-анализа с упором на то, что такое дипфейки, кто их производит, каковы преимущества и угрозы технологии дипфейков, некоторые текущие примеры дипфейков и способы борьбы с ними.

**ЧТО ТАКОЕ ДИПФЕЙКИ?**

Дипфейки, сочетающие в себе «глубокое обучение» и «фейк», представляют собой гиперреалистичные видеоролики, обработанные в цифровом формате, чтобы изображать людей, говорящих и делающих то, чего на самом деле никогда не происходило.

Дипфейки полагаются на нейронные сети, которые анализируют большие объемы образцов данных, чтобы научиться имитировать выражение лица, манеры, голос и интонации человека. Процесс включает в себя передачу видеозаписей двух людей в алгоритм глубокого обучения, чтобы научить его менять лица. Другими словами, дипфейки используют технологию картирования лица, а искусственный интеллект, заменяет лицо человека на видео лицом другого человека. Дипфейки стали известны широкой публике в 2017 году, когда пользователь Reddit опубликовал видеоролики, показывающие знаменитостей в компрометирующих сексуальных ситуациях. Дипфейки трудно обнаружить, поскольку они используют реальные кадры, могут иметь аутентичный звук и оптимизированы для быстрого распространения в социальных сетях. Таким образом, многие зрители предполагают, что видео, которое они смотрят, является подлинным.

Дипфейки нацелены на платформы социальных сетей, где легко распространяются заговоры, слухи и дезинформация, поскольку пользователи склонны идти с толпой.

В то же время продолжающийся «инфо апокалипсис», который заставляет людей думать, что они не могут доверять никакой информации, если она не исходит из их социальных сетей, включая членов семьи, близких друзей или родственников, и поддерживает мнения, которых они уже придерживаются. Фактически, многие люди открыты ко всему, что подтверждает их существующие взгляды, даже если они подозревают, что это может быть ложью.

Дешевые фейки (видеоролики низкого качества со слегка подделанным реальным содержанием) уже повсюду, поскольку широко доступно недорогое оборудование, такое как эффективные графические процессоры.

Программное обеспечение для создания высококачественных, реалистичных дипфейков для дезинформации все чаще становится доступным с открытым исходным кодом. Это позволяет пользователям, обладающими опосредованными техническими навыками без каких-либо художественных навыков почти идеально редактировать видео, менять лица, изменять выражения и синтезировать речь.

Что касается технологий, дипфейки - это продукт генеративно-состязательных сетей (GAN), а именно двух искусственных нейронных сетей, работающих вместе для создания реалистично выглядящих медиа. Эти две сети, так называемые «генератор» и «дискриминатор», обучаются на одном и том же наборе данных изображений, видео или звуков.

Затем первая пытается создать новые образцы, которые достаточно хороши, чтобы обмануть вторую сеть, которая определяет, является ли новый носитель реальным. Таким образом, они побуждают друг друга совершенствоваться. GAN может просмотреть тысячи фотографий человека и создать новый портрет, который приблизительно соответствует этим фотографиям, но не является точной копией ни одной из них.

В ближайшем будущем GAN будут обучаться меньшему количеству информации и смогут менять головы, целые тела и голоса. Хотя дипфейки обычно требуют большого количества изображений для создания реалистичной подделки, исследователи уже разработали метод создания фальшивого видео, загружая в него только одну фотографию, например, селфи.

**ПРЕИМУЩЕСТВА ТЕХНОЛОГИИ DEEPFAKE**

Технология Deepfake также находит положительное применение во многих отраслях, включая кино, образовательные средства массовой информации и цифровые коммуникации, игры и развлечения, социальные сети и здравоохранение, материаловедение и различные сферы бизнеса, такие как мода и электронная коммерция.

Киноиндустрия может извлечь выгоду из технологии дипфейков разными способами. Например, она может помочь в создании цифровых голосов для актеров, потерявших голоса из-за болезни, или для обновления отснятого материала вместо его пересъемки. Кинорежиссеры смогут воссоздавать классические сцены в фильмах, создавать новые фильмы с участием давно умерших актеров, использовать спецэффекты и продвинутый монтаж лиц в пост-продакшене, а также улучшать любительские видеоролики до профессионального качества.

Технология Deepfake также позволяет автоматически и реалистично озвучивать фильмы на любом языке, что позволяет разной аудитории лучше наслаждаться фильмами и образовательными материалами. Глобальная кампания по повышению осведомленности о малярии в 2019 году с участием Дэвида Бекхэма разрушила языковые барьеры с помощью образовательной рекламы, в которой использовались визуальные технологии и технологии изменения голоса, чтобы заставить его выглядеть многоязычным.

Аналогичным образом, технология дипфейков может преодолеть языковой барьер во время видеоконференций, переводя речь и одновременно изменяя движения лица и рта, чтобы улучшить зрительный контакт и создать впечатление, что все говорят на одном языке.

Технология, лежащая в основе дипфейков, позволяет создавать многопользовательские игры и миры виртуальных чатов с повышенным телеприсутствием, естественно звучащими и выглядящими умными помощниками, и цифровыми двойниками людей. Это помогает развивать лучшие человеческие отношения и взаимодействие в Интернете.

Также, технология может найти положительное применение в социальной и медицинской сферах. Дипфейки могут помочь людям справиться с потерей близких, «вернув к жизни» умершего друга в цифровом виде и тем самым потенциально помогая скорбящему близкому человеку попрощаться с ней.

Технология Deepfake может даже помочь людям с болезнью Альцгеймера взаимодействовать с более молодым лицом, которое они, возможно, помнят. Ученые также изучают возможность использования GAN для обнаружения аномалий в рентгеновских лучах и их потенциал в создании виртуальных химических молекул для материаловедения и медицинских открытий.

Компании заинтересованы в потенциале технологии дипфейков, применимой к брендам, поскольку она может существенно трансформировать электронную коммерцию и рекламу, например, бренды могут нанимать супермоделей, которые на самом деле супермоделями не являются, и демонстрировать модные наряды на различных моделях с разным оттенком кожи, ростом и весом.

Кроме того, дипфейки позволяют создавать сверхличный контент, который превращает самих потребителей в модели. Эта технология позволяет провести виртуальную примерку, чтобы предварительно просмотреть, как на них будет смотреться наряд, перед покупкой, а также может генерировать таргетированную модную рекламу, которая меняется в зависимости от времени, погоды и зрителя. Очевидным потенциальным применением является возможность быстро примерить одежду онлайн. Данная технология не только позволяет людям создавать цифровые клоны самих себя и путешествовать с этими личными аватарами по интернет-магазинам, но также примерять свадебное платье или костюм в цифровой форме, а затем виртуально ощущать место проведения свадьбы.

Кроме того, ИИ может создавать уникальные искусственные голоса, которые отличают компании и продукты и облегчают различение брендов.

**КТО ПРОИЗВОДИТ ДИПФЕЙКИ?**

Существует как минимум четыре основных типа производителей дипфейков:

1. Сообщества любителей дипфейков;
2. Политические игроки, такие как иностранные правительства и различные активисты;
3. Другие злонамеренные субъекты, такие как мошенники;
4. Законные субъекты, такие как телекомпании.

Людей в сообществах любителей дипфейков сложно выследить. После того, как в конце 2017 года один из пользователей разместил на Reddit дипфейки со знаменитостями, только что созданному сообществу любителей дипфейков понадобилось всего несколько месяцев, чтобы набрать 90 000 участников.

Многие любители сосредотачивают внимание на дипфейках, связанных с порнографией, в то время как другие помещают известных актеров в фильмы, в которых они никогда не создавали комических эффектов.

В целом любители склонны рассматривать видеоролики, созданные искусственным интеллектом, как новую форму онлайн-юмора и вклад в развитие таких технологий как решение интеллектуальной головоломки, а не как способ обмануть или запугать людей. Их дипфейки призваны быть развлекательными, забавными или политически сатирическими и могут помочь привлечь подписчиков в социальных сетях.

Некоторые любители могут искать более конкретные личные выгоды, такие как повышение осведомленности о потенциале технологии дипфейков, чтобы получить оплачиваемую работу, связанную с дипфейками, например, в музыкальных клипах или телешоу. Таким образом, любители и субъекты права, такие как телекомпании, могут сотрудничать друг с другом.

Хотя дипфейки в стиле мемов, созданные любителями, могут развлекать онлайн-пользователей, в них участвуют и более злонамеренные субъекты. Различные политические игроки, в том числе политические агитаторы, хактивисты, террористы и иностранные государства, могут использовать дипфейки в кампаниях по дезинформации и манипулирования общественным мнением и подрыва доверия к институтам той или иной страны.

В наши времена гибридной войны дипфейки представляют собой оружие дезинформации, направленное на вмешательство в выборы и сеяние гражданских беспорядков. Мы можем ожидать, что все больше и больше финансируемых государством и зарубежными интернет-«ферм троллей» будут использовать ИИ для создания и распространения политических фейковых видеороликов, адаптированных к конкретным предубеждениям пользователей социальных сетей.

Дипфейки также все чаще используются мошенниками с целью манипулирования рынком и акциями, а также различных других финансовых преступлений. Преступники уже использовали фальшивые аудиозаписи, созданные ИИ, чтобы выдать себя за руководителя, разговаривающего по телефону и просящего срочный денежный перевод. В будущем видеозвонки также можно будет имитировать в режиме реального времени. Визуальные материалы, необходимые для имитации руководителей, часто доступны в Интернете. Технология Deepfake может использовать визуальные и звуковые имитации руководителей, например, из видеороликов TED Talk, доступных на YouTube.

**ВОЗМОЖНЫЕ УГРОЗЫ ДИПФЕЙКОВ**

Дипфейки представляют собой серьезную угрозу нашему обществу, политической системе и бизнесу, поскольку они:

1. Оказывают давление на журналистов, пытающихся отделить настоящие новости от фейковых;
2. Угрожают национальной безопасности России, распространяя пропаганду и вмешиваясь в выборы;
3. Подрывают доверие граждан к информации властей;
4. Поднимать вопросы кибербезопасности для людей и организаций.

Весьма вероятно, что журналистской индустрии придется столкнуться с серьезной проблемой доверия потребителей из-за дипфейков.

Дипфейки представляют б**О**льшую угрозу, чем «традиционные» фейковые новости, потому что их сложнее обнаружить, и люди склонны верить, что фейк реален. Эта технология позволяет создавать, казалось бы, законные новостные видеоролики, которые ставят под угрозу репутацию журналистов и средств массовой информации.

Кроме того, победа в гонке за доступ к видеозаписям, снятым свидетелем инцидента, может обеспечить конкурентное преимущество новостному средству массовой информации, в то время как опасность возрастает, если предлагаемые кадры являются фейковыми. Во время всплеска напряженности между Индией и Пакистаном в 2019 году агентство «Reuters» обнаружило 30 фейковых видеороликов об инциденте; в основном старые видео с других мероприятий публикуются с новыми субтитрами. Неправильно приписываемые видеоматериалы, такие как реальный марш протеста или жестокая стычка, снабженные подписями, намекающими на то, что это произошло где-то в другом месте, являются растущей проблемой, и она будет усугубляться ростом количества дипфейков.

В поисках видео очевидцев массовой стрельбы в Крайстчерче, Новая Зеландия, агентство «Reuters» наткнулось на видео, на котором, как утверждается, запечатлен момент, когда подозреваемый был застрелен полицией. Однако они быстро обнаружили, что это был другой инцидент в США, и подозреваемый в стрельбе в Крайстчерче не был убит.

Разведывательное сообщество многих стран обеспокоено тем, что дипфейки будут использоваться для угрозы национальной безопасности своих государств путем распространения политической пропаганды и срыва избирательных кампаний. Представители американской разведки, например, неоднократно предупреждали об угрозе иностранного вмешательства в американскую политику, особенно в преддверии выборов.

Вложить слова в чьи-то уста видео, которое становится вирусным, является мощным оружием в сегодняшних дезинформационных войнах, поскольку такие измененные видео могут легко исказить мнение избирателей. Любая разведка мира может создать дипфейковое видео, на котором политик использует расовый эпитет или берет взятку, кандидат в президенты признается в соучастии в преступлении или предупреждает другую страну о предстоящей войне, правительственный чиновник в, казалось бы, компрометирующей ситуации или признание секретного плана по осуществлению заговора или признание солдатами военных преступлений, таких как убийства мирных жителей за рубежом. Хотя такие фальшивые видеоролики, скорее всего, вызовут внутренние волнения, беспорядки и срывы выборов, другие государства могут даже решить проводить свою внешнюю политику, основанную на нереальности, что приведет к международным конфликтам.

Дипфейки, вероятно, будут препятствовать цифровой грамотности и доверию граждан к информации, предоставляемой властями, поскольку фейковые видео, на которых правительственные чиновники говорят о вещах, которых никогда не было, заставляют людей сомневаться в их авторитетности. Действительно, сегодня люди все чаще страдают от спама, генерируемого искусственным интеллектом, а также от фейковых новостей, основанных на фанатичных текстах, фальшивых видеороликах и множестве теорий заговора.

Тем не менее, наиболее разрушительным аспектом дипфейков, возможно, является не дезинформация сама по себе, а то, как постоянный контакт с дезинформацией заставляет людей чувствовать, что большому количеству информации, включая видео, просто нельзя доверять, что приводит к явлению, известному как «информационный апокалипсис» или «информационный апокалипсис», «апатия к реальности».

Более того, люди могут даже отвергнуть подлинные кадры как фейковые просто потому, что они укоренились в представлении, что все, во что они не хотят верить, должно быть фейковым. Другими словами, самая большая угроза заключается не в том, что люди будут обмануты, а в том, что они начнут считать все обманом.

Проблемы кибербезопасности представляют собой еще одну угрозу, исходящую от дипфейков. Корпоративный мир уже выразил заинтересованность в защите от вирусного мошенничества, поскольку дипфейки могут использоваться для манипулирования рынком и акциями, например, показывая, как генеральный директор произносит расистские или женоненавистнические оскорбления, объявляет о фиктивном слиянии, делает ложные заявления о финансовых потерях или банкротства, или изображая их как совершивших преступление.

Кроме того, фейковая порнография или анонсы продуктов могут быть использованы для саботажа бренда, шантажа или для того, чтобы поставить в неловкое положение руководство. Технология deepfake позволяет в режиме реального времени выдавать себя за руководителя в цифровом формате, например, попросить сотрудника выполнить срочный перевод денег или предоставить конфиденциальную информацию, может создать мошенническую личность и в прямых трансляциях превратить лицо взрослого человека в лицо ребенка или молодого человека, вызывая обеспокоенность по поводу использования этой технологии. Наконец, дипфейки могут способствовать распространению вредоносных скриптов.

Недавно исследователи обнаружили, что веб-сайт, посвященный дипфейкам, использовал компьютеры своих посетителей для майнинга криптовалют. Как итог, любители дипфейков могут стать объектами «криптоджекинга», поскольку у них, вероятно, есть мощные компьютеры.

**ПРИМЕРЫ ДИПФЕЙКОВ**

Большинство дипфейков сегодня на социальных платформах, таких как YouTube или Facebook[[2]](#footnote-2), можно рассматривать как безобидное развлечение или художественные произведения с использованием мертвых или живых общественных деятелей. Но, есть и примеры обратной - темной стороны дипфейков, а именно порноместь, а также попытки политического и неполитического влияния.

Многие дипфейки фокусируются на знаменитостях, политиках и корпоративных лидерах, потому что в Интернете полно исходных фотографий и видео с ними, на основе которых можно создать большие объёмы изображений, необходимые для обучения системы дипфейков искусственного интеллекта. Большинство таких дипфейков - это шутки, розыгрыши и забавные мемы с комедийным или сатирическим эффектом.

Дипфейк может показать, например, что Николас Кейдж играет в фильмах, в которых он никогда не снимался, таких как «Индиана Джонс» или «Терминатор 2». Некоторые интригующие примеры дипфейков включают видео, в котором Олден Эренрайк заменяется молодым Харрисоном Фордом в клипах из фильма «Соло: Истории звездных войн», а также видео актера Билла Хейдера, появляющегося в сериале «Поздно вечером» с Дэвидом Леттерманом.

Пока Хейдер говорит о Томе Крузе, его лицо превращается в лицо Круза. Некоторые дипфейки показывают мертвых знаменитостей, таких как лицо бывшего вокалиста группы «Queen» Фредди Меркьюри, наложенное на лицо актера Рами Малека, а также русский мистик Григорий Распутин, поющий мощную балладу Бейонсе «Halo».

Художественный музей в США использовал эту технологию, чтобы вернуть к жизни Сальвадора Дали, чтобы приветствовать посетителей, а другая система искусственного интеллекта заставляет любого танцевать, как прима-балерина, навязывая движения настоящего танцора телу целевого человека, тем самым создавая видео, на котором цель показана как профессиональная танцовщица.

Однако, примеры вредоносных дипфейков появляются все чаще и чаще. Технология Deepfake позволяет создавать порноместь, то есть принудительную порнографию с использованием изображений обычных людей, которые публикуются в социальных сетях без их согласия. Так, такие знаменитости, как Скарлетт Йоханссон, появлялись в фейковых фильмах для взрослых, в которых их лица накладывались на лица порнозвезд. На политической сцене в дипфейке 2018 года, созданном голливудским режиссером Джорданом Пилом, бывший президент США Обама обсуждал опасности фейковых новостей и высмеивал нынешнего президента Трампа.

В 2019 году измененное видео американского политика Нэнси Пелоси стало вирусным и имело широкий охват; видео было замедлено, чтобы она выглядела пьяной. В фейковом видео 2018 года Дональд Трамп дал совет народу Бельгии по поводу изменения климата. Видео было создано бельгийской политической партией «sp.a» с целью привлечь людей подписать онлайн-петицию, призывающую правительство Бельгии принять более срочные меры по борьбе с изменением климата. Видео вызвало возмущение по поводу вмешательства американского президента в чужую страну в климатическую политику Бельгии. В 2019 году Демократическая партия США сфальсифицировала своего председателя Тома Переса, чтобы подчеркнуть потенциальную угрозу дипфейков выборам 2020 года.

Хотя это примеры ограниченного политического влияния, другие дипфейки могут иметь более продолжительное воздействие. В Центральной Африке фейк с изображением давно не появлявшегося на публике президента Габона Али Бонго (как полагали, был слаб здоровьем или умер) был назван спусковым крючком для неудачного государственного переворота, совершенного габонскими военными. В Малайзии политический скандал вызвал вирусный клип с дипфейком, в котором мужчина признался в сексе с министром. Также для создания дипфейков использовались лица, не являющиеся политическими. В июне 2022 года высококачественный дипфейк двух британских художников с участием генерального директора Facebook [[3]](#footnote-3)Марка Цукерберга собрал миллионы просмотров. Видео ложно изображает Цукерберга, отдающего дань уважения вымышленной злой организации «Спектр» из сериала о Джеймсе Бонде, которая учит его, как получить полный контроль над конфиденциальными данными миллиардов людей и, таким образом, владеть их будущим. Целью видео было использование новостных материалов, технологии дипфейков и актера озвучивания, чтобы показать, как можно использовать технологии для манипулирования данными.

**МЕТОДЫ БОРЬБЫ С ДИПФЕЙКАМИ**

Рассмотренные новостные статьи предполагают, что существует четыре способа борьбы с дипфейками:

1. Законодательство и регулирование;
2. Корпоративная политика и добровольные действия;
3. Образование и обучение;
4. Технология борьбы с дипфейками, которая включает обнаружение дипфейков, аутентификацию контента и предотвращение дипфейков.

Законодательство и регулирование являются очевидными средствами борьбы с дипфейками. В настоящее время дипфейки конкретно не регулируются гражданским или уголовным законодательством, хотя эксперты по правовым вопросам предлагают адаптировать действующие законы для охвата клеветы, диффамации, мошенничества с личными данными или выдачи себя за государственного чиновника с использованием дипфейков.

Растущая сложность технологий искусственного интеллекта требует принятия новых типов законов и нормативной базы. Например, дипфейки вызывают обеспокоенность по поводу конфиденциальности и авторских прав, поскольку визуальные изображения людей в дипфейковых видео не являются точными копиями какого-либо существующего материала, а скорее новыми изображениями, созданными ИИ. Таким образом, регулирующие органы должны ориентироваться в сложном правовом ландшафте вокруг законов о свободе слова и собственности, чтобы должным образом регулировать использование технологии дипфейков.

С другой стороны, подходящим правовым решением проблемы распространения вредных дипфейков не был бы полный запрет этой технологии, что было бы неэтично. Хотя новые законы могут быть приняты для предотвращения дипфейков, они также нуждаются в механизмах обеспечения соблюдения.

Сегодняшние компании, занимающиеся социальными сетями, пользуются широким иммунитетом в отношении контента, который пользователи публикуют на своих платформах. Одним из законодательных вариантов может стать лишение компаний социальных сетей юридического иммунитета от контента, публикуемого их пользователями, что сделает не только пользователей, но и платформы более ответственными за публикуемые материалы. Тем не менее, законодательство мало повлияло на злонамеренных субъектов, таких как иностранные государства и террористы, которые могут проводить массовые кампании дезинформации против других государств в социальных сетях.

Корпоративная политика и добровольные действия могут стать более эффективными инструментами против дипфейков. Например, политики могут взять на себя обязательство не использовать незаконные тактики цифровых кампаний и не распространять дезинформацию, такую ​​как дипфейки, в своих избирательных кампаниях.

Компании, занимающиеся социальными сетями, должны обеспечить соблюдение этических норм и отказаться от того факта, что размещение вызывающего разногласия контента в верхней части ленты является финансовой выгодой, поскольку это максимизирует время взаимодействия с рекламой. Немногие компании, занимающиеся социальными сетями, пока имеют политику в отношении дипфейков, им следует сотрудничать, чтобы не допустить использования своих платформ в качестве оружия для дезинформации, и, таким образом, активно применять прозрачные общие политики для блокировки и удаления дипфейков. Сейчас многие компании не удаляют спорный контент, а понижают его рейтинг, чтобы его было труднее найти, поскольку они становятся менее заметными в новостных лентах пользователей. С другой стороны, рост разжигания ненависти, фейковых новостей и дезинформации, загрязняющей цифровые платформы, побудил некоторые компании принять дополнительные меры, такие как приостановка учетных записей пользователей и инвестиции в технологии более быстрого обнаружения.

«Reddit[[4]](#footnote-4)» и «Pornhub[[5]](#footnote-5)» запретили дипфейк-порно и другую порнографию без согласия и принимают меры в случае, если пользователи отмечают такие материалы. Facebook отсекает любой контент, который сторонние специалисты по проверке фактов считают ложным или вводящим в заблуждение, от показа рекламы и зарабатывания денег; компания работает с более чем 50 организациями по проверке фактов, учеными, экспертами и политиками, чтобы найти новые решения.

Алгоритмы «Instagram[[6]](#footnote-6)» не будут рекомендовать людям просматривать контент, помеченный службами проверки фактов как «ложный». Среди новостных медиа-компаний «Wall Street Journal» и «Reuters» сформировали корпоративные группы, которые помогают и обучают своих репортеров выявлять фейковый контент, а также внедрять методы и инструменты обнаружения, такие как перекрестные ссылки на местоположение на картах Google и обратный поиск изображений.

Образование и подготовка имеют решающее значение для борьбы с дипфейками. Несмотря на широкое освещение в новостях и обеспокоенность властей, общественность еще не приняла во внимание угрозу дипфейков. В целом существует необходимость повышения осведомленности общественности о возможности злоупотребления ИИ. В то время как дипфейки предоставляют киберпреступникам новые инструменты для социальной инженерии, компаниям и организациям необходимо быть в состоянии повышенной готовности и разработать планы киберустойчивости. Правительства, регулирующие органы и частные лица должны понимать, что видео, вопреки внешнему виду, может не давать точного представления о том, что произошло, и знать, какие перцептивные сигналы могут помочь идентифицировать дипфейки. В школах рекомендуется обучать критическому мышлению и цифровой грамотности, поскольку эти качества способствуют способности детей распознавать фейковые новости и более уважительно взаимодействовать друг с другом в Интернете.

Эти навыки также следует пропагандировать среди старшего, менее технически подкованного населения. Причина этого в том, что люди должны иметь возможность критически оценить подлинность и социальный контекст видео, которое они, возможно, захотят просмотреть, а также надежность его источника (то есть того, кто поделился видео и что в нем говорится), чтобы понять истинную цель видео.

Также важно помнить, что качество не является показателем подлинности видео. Кроме того, люди должны понимать, что по мере развития технологий для создания дипфейков будет требоваться все меньше фотографий реальных лиц и что никто от этого не застрахован.

Любой, кто публикует в социальной сети одно селфи или видео со скоростью 30 кадров в секунду, рискует стать дипфейком. Хотя лучший метод — не допускать попадания фотографий и видео в Интернет, некоторую защиту могут обеспечить такие препятствия, как махание рукой перед лицом на фотографии или видео.

Компании, правительства и органы власти, использующие технологию распознавания лиц и хранящие огромные объемы данных о лицах в целях безопасности и проверки, должны устранить угрозу кражи личных данных в случае утечки таких данных.

Технология защиты от дипфейков предоставляет, пожалуй, самый разнообразный набор инструментов для:

1. Обнаружения дипфейков;
2. Аутентификации контента;
3. Предотвращения использования контента для создания дипфейков.

В целом, проблемы технологий аутентификации контента и выявления фейков связаны с их масштабностью, а также с тем фактом, что существует гораздо больше доступных исследовательских ресурсов и людей, работающих над разработкой технологий для создания дипфейков, чем над технологиями их обнаружения (например, пользователи загружают на YouTube [[7]](#footnote-7)500 часов контента в минуту. Твиттер борется с 8 миллионами аккаунтов в неделю, которые пытаются распространять контент с помощью манипулятивной тактики). Это создает огромные проблемы для технологий, чтобы пройти весь опубликованный материал за короткое время. Кроме того, разработчики дипфейков склонны использовать результаты опубликованных исследований дипфейков для совершенствования своих технологий и обхода новых систем обнаружения (например, исследователи обнаружили, что ранние методы дипфейка не могли имитировать скорость моргания человека; тогда как недавние программы зафиксировали отсутствие моргания или неестественное моргание после публикации результатов).

Если брать за пример США, там финансирование разработки по обнаружению дипфейков в основном поступает от агентств национальной безопасности, таких как Агентство перспективных исследовательских проектов Министерства обороны США (DARPA), у частных компаний, занимающихся кибербезопасностью, существуют значительные бизнес-возможности для разработки решений для обнаружения дипфейков, создания надежных платформ, борьбы с незаконными ботами и борьба с мошенничеством и цифровым загрязнением. Однако одного только развития технологии борьбы с дипфейками недостаточно. Организации также должны внедрить эти технологии; например, правительство в любой конкретной стране может быть модернизировано, чтобы противостоять дипфейкам и помогать защищать своих граждан от дипфейков.

Эксперты по медиа-криминалистике предложили тонкие индикаторы для обнаружения дипфейков, включая ряд недостатков, таких как дрожание лица, мерцание и искажение; волнистость в движениях человека; несогласованность речи и движений рта; ненормальные движения неподвижных объектов, таких как стойка микрофона; несоответствия освещения, отражений и теней; размытые края; ракурсы и размытость черт лица; отсутствие дыхания; неестественное направление глаз; отсутствуют черты лица, такие как известная родинка на щеке; мягкость и вес одежды и волос; слишком гладкая кожа; отсутствуют детали волос и зубов; нарушение симметрии лица; несоответствия уровней пикселей; и странное поведение человека, делающего что-то неправдоподобное.

Людям становится все труднее отличить настоящее видео от подделки, искусственный интеллект может сыграть важную роль в обнаружении дипфейков (пример: алгоритмы искусственного интеллекта могут анализировать закономерности неравномерности фотоотклика в отснятом материале, то есть недостатки, уникальные для датчика освещенности конкретных моделей камер, или биометрические данные, такие как кровоток, на которые указывают тонкие изменения, происходящие на лице человека. в видео).

Новые алгоритмы обнаружения фейков основаны на слуховых системах млекопитающих, например, на том, как мыши обнаруживают несоответствия и тонкие ошибки в звуке, которые часто игнорируются людьми. ИИ может либо просматривать видео покадрово, чтобы отслеживать признаки подделки, либо просматривать все видео сразу, чтобы изучить мягкие биометрические сигнатуры, включая несоответствия в аутентифицированных отношениях между движениями головы, моделями речи и выражениями лица, такими как улыбаясь, чтобы определить, не были ли манипуляции с видео. Последний метод можно адаптировать для отдельных лиц, например, для высокопоставленных политиков, которых, скорее всего, подделывают.

Проблема с дипфейками заключается не только в том, чтобы доказать, что что-то является ложным, но и в том, чтобы доказать подлинность объекта. Аутентификация видео особенно важна для компаний СМИ, которым приходится определять подлинность видео, распространяемого в ненадежной среде, в которой сложно отследить детали создателя, происхождения и распространения видео.

Предлагаемые решения для аутентификации контента варьируются от цифровых водяных знаков до методов цифровой криминалистики. Было бы идеально создать некий «фильтр правды», автоматизированную систему в Интернете, которая бы обеспечивала поддельные и подлинные измерения видео; таким образом, каждое видео, публикуемое на сайте социальной сети, будет проходить процесс аутентификации (например, программное обеспечение, встроенное в камеры смартфонов, может создать цифровой отпечаток пальца в момент записи фильма. При воспроизведении отснятого материала его водяной знак можно сравнить с исходным отпечатком пальца, чтобы проверить совпадение и предоставить зрителю оценку, указывающую на вероятность подделки).

Действительно, цифровые водяные знаки, такие как хеширование, могут предоставить видеофайлу короткую строку чисел, которая теряется при манипуляциях с видео. Он также может предоставить подтвержденное алиби общественным деятелям, поскольку они постоянно записывают, где они находятся и что делают. Подлинность видео также обеспечивается путем сопоставления его происхождения, то есть того, было ли видео изначально получено из авторитетного источника и как оно с тех пор распространилось в Интернете. Технология блокчейн может помочь в проверке происхождения и распространения видео путем создания и хранения цифровых подписей в реестре, манипулировать которым практически невозможно. Платформы социальных сетей, новостные компании и другие онлайн-участники должны затем продвигать видео, подлинность которых подтверждена, по сравнению с непроверенными видео. Тем не менее, всегда найдутся люди, которые решат не верить инструменту проверки и, скорее, по-прежнему будут иметь желание потреблять и поддерживать фейковые СМИ.

Наконец, технологии могут предотвратить создание дипфейков путем вставки «шума» в фотографии или видео. Этот шум незаметен для человеческого глаза, но не позволяет использовать визуальный материал в автоматизированном программном обеспечении для дипфейков.

Можно также носить специально разработанные очки, напечатанные на 3D-принтере, чтобы избежать распознавания лиц, обманом заставив программное обеспечение Deepfake неправильно классифицировать пользователя. Эта технология может помочь вероятным целям, таким как политики, знаменитости и руководители, предотвратить создание из них дипфейков. Кроме того, исследователи, разрабатывающие технологии GAN, могут разработать и внедрить надлежащие меры защиты, чтобы их технологии было сложнее использовать в целях дезинформации.

Как и в сфере кибербезопасности в целом, первым шагом к решению является понимание проблемы и ее способности влиять на нас. Только тогда становится возможным разработать и реализовать технические решения, способные решить поставленные задачи. Тем не менее, ни одно из технологических решений не может полностью устранить риск дипфейков, а технологический подход (каждая проблема имеет технологическое решение) может даже дезориентировать дискуссию от более экзистенциальных вопросов о том, почему существуют дипфейки и какие другие угрозы может создать ИИ. Поэтому наиболее эффективные способы борьбы с распространением дипфейков включают сочетание юридических, образовательных и социотехнических достижений.

**ВЫВОД**

В ходе исследования были рассмотрены и проанализированы открытые новостные статьи о дипфейках, чтобы лучше понять, что такое дипфейки и кто их производит, преимущества и угрозы технологии дипфейков, примеры современных дипфейков и способы борьбы с ними.

При этом моё исследование показало, что дипфейки - это гиперреалистичные видеоролики, обработанные в цифровом формате, чтобы изобразить людей, говорящих и делающих то, чего никогда в реальности не происходило.

Дипфейки создаются с использованием искусственного интеллекта, то есть генеративно-состязательных сетей (GAN), которые противопоставляют дискриминационные и генеративные алгоритмы друг другу для точной настройки производительности при каждом повторении и, таким образом, создают убедительную фальшивку. Эти фейки реальных людей зачастую очень вирусные и имеют тенденцию быстрого распространения через платформы социальных сетей, что делает их эффективным инструментом дезинформации.

Результаты этого краткого исследования возможно внесут определенный вклад в формирующуюся научную литературу по дипфейкам. Мои результаты подтверждают указывают, что коммерческие платформы состоят как из платформ новостных СМИ, ряда платформ социальных сетей, и дипфейки подпитываются не только горячим политическим, но и текущим социальным контекстом из-за так называемого информационного апокалипсиса, который заставляет людей перестать доверять информации, исходящей не из их личных социальных сетей и не соответствующей их прежним убеждениям.

Рост числа бизнес-моделей фейковых новостей, которые генерируют веб-трафик на страницы фейковых новостей для получения прибыли за счет рекламы только увеличивается.

Мотивация действий многих субъектов может включать злонамеренное причинение вреда другим людям различными способами, манипулирование общественным мнением по конкретным темам, посеять замешательство или разногласия в обществе, извлечь финансовую выгоду или банальное развлечение. Безусловно, существуют лица и организации, такие как телекомпании, которые создают и поддерживают дипфейки с целью разработки и применения технологии дипфейков для законного использования, например, для оплачиваемой работы над музыкальными клипами.

Что касается законного использования технологий глубокого обучения, Netflix, например, может позволить зрителям выбирать актеров на экране перед тем, как начать воспроизведение, или даже позволить зрителям самим сниматься в фильме. Это исследование выявило ряд дополнительных областей законного использования технологии, включая образовательные средства массовой информации и цифровые коммуникации, игры и развлечения, социальную сферу и здравоохранение, материаловедение, а также различные области бизнеса, такие как мода и персонализированная электронная коммерция.

Согласно исследования, дипфейки представляют собой серьезную угрозу для общества, политической системы и бизнеса, поскольку они оказывают давление на журналистов, пытающихся отделить реальные новости от фейковых, угрожают национальной безопасности, распространяя пропаганду, которая вмешивается в выборы, подрывают доверие граждан к информации властей. и поднимать вопросы кибербезопасности для людей и организаций.

С другой стороны, существует как минимум четыре известных способа борьбы с дипфейками, а именно:

1. Законодательство и регулирование;
2. Корпоративная политика и добровольные действия;
3. Образование и обучение
4. Технологии борьбы с дипфейками.

Хотя законодательные меры могут быть приняты против некоторых производителей дипфейков, они не эффективны в отношении иностранных государств. Скорее, корпоративные политики и добровольные действия, такие как политика модерации контента, направленная против дипфейков, и быстрое удаление помеченного пользователями контента на платформах социальных сетей, а также образование и обучение, направленные на повышение грамотности в области цифровых медиа, улучшение поведения в Интернете и критического мышления, которые создают когнитивные и конкретные меры защиты от потребления и злоупотребления цифровым контентом, вероятно, будут более эффективными.

Государственным органам, компаниям, преподавателям и журналистам необходимо повысить осведомленность граждан об угрозах, которые ИИ представляет для доверия СМИ, и запретить мошенническое использование таких технологий в коммерческих, политических или антисоциальных целях.

Технологические решения, в том числе автоматизированные инструменты для обнаружения дипфейков, аутентификации контента и предотвращения дипфейков, представляют собой динамичную область методов безопасности. Следовательно, перед технологическими предпринимателями открываются многочисленные возможности для бизнеса, особенно в области кибербезопасности и искусственного интеллекта. Технология дипфейков развивается все более быстрыми темпами. Для людей быстро становится невозможным отличить настоящие видео от фейковых. Таким образом, наши результаты перечисляют многочисленные признаки обнаружения дипфейков и предлагают использовать ИИ для обнаружения фейков, созданных ИИ, в качестве эффективной стратегии борьбы.

Растет потребность в онлайн-проверке источников и аутентификации контента, и следует использовать повсеместные уровни проверки, основанные на цифровых водяных знаках. Может помочь еще одна новая технология, а именно блокчейн. Технология блокчейн не только обладает высокой устойчивостью к подделкам и может хранить данные доступным, безопасным, прозрачным и отслеживаемым способом, но также может отслеживать и сертифицировать происхождение и историю данных.

В духе обзорного исследования показанные результаты могут внести вклад в развивающуюся область дипфейков, объединяя разрозненные результаты редких академических исследований фейковых новостей и дипфейков, а также дополняя эти результаты примерами и обсуждением дипфейков, происходящих публично и в новостных статьях.

В исследовании, безусловно, есть ограничения. Во-первых, хотя эмпирическое исследование охватило новостные онлайн-статьи о дипфейках, в открытых источниках доступно гораздо больше статей. Учитывая скорость развития этой технологии, эти статьи также могут предоставить дополнительную информацию о дипфейках и предложить дополнительные методы борьбы с ними. Во-вторых, мой эмпирический материал был собран из открытых источников, а именно из новостных интернет-сайтов. Использование других типов данных, таких как дискуссии в онлайн-сообществах, посвященных дипфейкам, и интервью с разработчиками GAN и художниками дипфейков, некоторые из которых известны публике благодаря своей работе не только как разработчики технологий дипфейков, но и как разработчики технологий борьбы с дипфейками, может дать дополнительную информацию о стратегиях борьбы с дипфейками.

Разделы комментариев в некоторых проанализированных новостных статьях содержали большое количество мнений и идей читателей. Анализ этих комментариев может дать дополнительную информацию о том, как дипфейки воспринимаются более широкой аудиторией и, следовательно, на какие методы борьбы следует обратить внимание. Это открывает широкие возможности для будущих исследований дипфейков.

1. 21 марта 2022 г. компания Meta была признана экстремистской организацией на территории РФ. [↑](#footnote-ref-1)
2. 21 марта 2022 г. компания Meta была признана экстремистской организацией на территории РФ. [↑](#footnote-ref-2)
3. 21 марта 2022 г. компания Meta была признана экстремистской организацией на территории РФ. [↑](#footnote-ref-3)
4. Reddit – это англоязычная платформа с полезным контентом, которая сочетает в себе и соцсеть, и форум, и портал одновременно. [↑](#footnote-ref-4)
5. Pornhub — порносайт, предоставляющий доступ к порнографическим материалам. Является одним из самых популярных порнографических сайтов в мире. [↑](#footnote-ref-5)
6. 21 марта 2022 г. компания Meta была признана экстремистской организацией на территории РФ. [↑](#footnote-ref-6)
7. Канал признан экстремистским на территории РФ. [↑](#footnote-ref-7)