ЦЕНТР ГЛОБАЛЬНОЙ ИТ-КООПЕРАЦИИ

Аналитический обзор

Дипфейки в цифровом пространстве: основные международные подходы к исследованию и регулированию

Москва, 2023 г.

АНО «Центр компетенций глобальной ИТ-кооперации»
ANO «Center for Global IT-Cooperation»
https://cgitc.ru

Авторы:
Игнатьев А.Г., руководитель аналитического направления CGITC
Курбатова Т.А., старший аналитик CGITC
Корректор: Нагель Е.В.

Назначение исследования

Информация базируется на дискуссиях, научных работах, аналитических исследованиях, обзорных публикациях и нормативно-правовых актах в зарубежных странах.

Цель работы — представить краткий обзор зарубежных подходов для получения базовых представлений и выявления магистральных тенденций в этой сфере. На основе собранных данных могут быть рассмотрены возможности совершенствования инструментов регулирования дипфейков (ДФ) в России.

Материал может быть учтен в составе других экспертных документов при формировании решения о необходимости разработки в стране дополнительных нормативно-правовых актов. Обзор может быть использован также для различных текущих управленческих решений в области искусственного интеллекта (ИИ).

Важно отметить, что при выработке такого рода решений и регуляторных мер должны быть комплексно использованы фактические научные и исследовательские материалы, статистические данные и практический опыт, накопленный в профильных ведомствах и организациях, отвечающих за вопросы безопасности в Интернете и информационном пространстве в целом.

Опыт CGITC в исследовании проблемы

В 2022 году Центр глобальной ИТ-кооперации (CGITC) прошел конкурсный отбор и получил право участия в проекте Think20 (T20) с аналитическим обзором «Дипфейки и безопасность в информационной среде: вызовы для правительств, общества и бизнеса». Работа проводилась в целевой группе «Полноценное цифровое подключение, кибербезопасность и расширение возможностей» («Task Force 2 – Meaningful Digital Connectivity, Cybersecurity and Empowerment») в рамках тематики «Риски и угрозы для кибербезопасности и защита персональных данных» («Cybersecurity Risks, Threats and Data Privacy»).

T20 представляет собой международную исследовательскую сеть G20, которая работает как банк идей и исследований для подготовки документов и резолюций «Группы двадцати». В T20 участвуют исследовательские и научные организации, которые подают заявку и проходят конкурсный отбор, чтобы получить право на подготовку обзора по определенной теме – Policy Brief.

Аналитический обзор (Policy Brief) CGITC был опубликован на официальном сайте Think20[1] в сентябре 2022 года к саммиту T20 в Индонезии, который проходил с 4 по 6 сентября.

О Центре

АНО «Центр компетенций по глобальной ИТ-кооперации» создан в 2020 году для экспертного изучения вопросов международного сотрудничества в сфере информационных технологий (ИТ), укрепления позиций России в глобальной ИТ-кооперации, а также продвижения новых подходов к многостороннему управлению Интернетом.

CGITC является членом Сектора развития телекоммуникаций (ITU-D) Международного союза электросвязи, участником международного Форума по управлению интернетом (IGF), соорганизатором ежегодного Российского форума по управлению Интернетом.

Центр проводит исследования и реализует проекты в области цифровой грамотности, управления Интернетом, научно-технического сотрудничества в сфере цифровой экономики, оказывает практическое содействие новым командам и начинающим экспертам по продвижению инноваций и стартапов. Во взаимодействии с международным сообществом и при поддержке заинтересованных специалистов в России CGITC на регулярной основе проводит ряд научных и экспертных круглых столов, конференций и вебинаров.

ДОКУМЕНТ ДОСТУПЕН ПО ССЫЛКЕ:

www.t20indonesia.org/wp-content/uploads/2022/11/TF2_Deepfakes-and-Security-in-the-Information-Environment_Challenges-for-Governments-Society-and-Business-1.pdf

[1] www.t20indonesia.org

Правила использования обзора

Настоящий аналитический обзор «Дипфейки в цифровом пространстве: основные международные подходы к исследованию и регулированию» (далее — «обзор») подготовлен специалистами АНО «Центр глобальной ИТ-кооперации».

Информация, приведенная в обзоре, подпадает под действие Закона об авторских правах Российской Федерации. Исключительные права на обзор принадлежат АНО «Центр глобальной ИТ-кооперации» (далее — «правообладатель»).

Обзор может использоваться в целях ознакомления. Допускается размещение активных ссылок на него в информационных источниках без непосредственного копирования его содержания. При любом использовании обзора активная ссылка на источник обязательна.

Частичное или полное воспроизведение и распространение, а также любое коммерческое использование обзора запрещено без письменного разрешения правообладателя, а также без ссылки на авторов исследования.

Приступая к ознакомлению с обзором, вы подтверждаете свое согласие с изложенными ниже условиями:

- Правообладатель не принимает на себя обязательства или ответственность за использование информации, содержащейся в обзоре.
- Обзор носит исключительно информационный характер и составлен на основе открытых источников, признанных надежными, однако правообладатель не несет ответственность за точность приведенных данных.
- Выводы, представленные в обзоре, также носят исключительно информационный характер и основаны на данных, полученных из открытых источников, указанных в сносках к обзору.
- Обзор не является юридическим заключением по вопросам, рассмотренным в нем. Правообладатель не несет ответственность за решения, принятые на основании представленных в обзоре данных.
- Обзор также включает в себя ссылки на сторонние веб-сайты, находящиеся вне контроля правообладателя. Правообладатель не несет ответственность за содержание этих ссылок. Такая ответственность во всех случаях возлагается на соответствующего провайдера либо оператора этих сторонних веб-сайтов.

Оглавление

1. АКТУАЛЬНОСТЬ ПРОБЛЕМАТИКИ И ОБЩЕЕ СОСТОЯНИЕ ИССЛЕДОВАНИЙ 7

2. ОБЩАЯ КЛАССИФИКАЦИЯ ДИПФЕЙКОВ 10

2.1. Использование термина **10**
2.2. Признаки и критерии, отличающие дипфейки от иных видов фейков **12**
2.2.1. Дешевые (cheapfakes) и мелкие (shallowfakes) подделки **12**
2.2.2. Типы фейков в Интернете **14**
2.2.3. Источники фейковых новостей **15**
2.3. Общая классификация дипфейков **16**
2.3.1. По методам создания **16**
2.3.2. По создателям **17**
2.3.3. Полезные или безвредные дипфейки **18**
2.3.4. Вредоносные дипфейки **20**
2.3.4.1. По типам злонамеренных и преступных целей **20**
2.3.4.2. По наносимому ущербу **21**

3. ВЫЗОВЫ И НАИБОЛЕЕ ПРОБЛЕМНЫЕ ВОПРОСЫ, СВЯЗАННЫЕ С ДИПФЕЙКАМИ 22

3.1. Угрозы **23**
3.1.1. Политика **23**
3.1.2. Судебная система **24**
3.1.3. Социальная инженерия **24**
3.1.4. Финансовая сфера **25**
3.1.5. Бизнес **25**
3.1.6. Цифровые платформы и социальные сети **26**
3.1.7. Дети **26**
3.1.8. Доверие к информации **26**
3.1.9. Дезинформация **27**
3.2. Отдельные технические вопросы, связанные с дипфейками, некоторые аспекты обнаружения ДФ **28**

4. НАИБОЛЕЕ ЗНАЧИМЫЕ ЗАРУБЕЖНЫЕ ПРАКТИКИ И ПОДХОДЫ 30

4.1. Исследования и инициативы **30**
4.2. Внутрикорпоративная политика и документы **35**
4.2.1. Twitter **35**
4.2.2. Meta* (Facebook*, Instagram*) **35**
4.2.3. Snapchat/TikTok **36**
4.2.4. YouTube **37**
4.2.5. Microsoft **37**
4.3. Рассматриваемые и принятые нормативно-правовые акты, другие инструменты регулирования, правоприменительные практики **38**
4.3.1. Европейский союз **38**
4.3.2. Республика Индия **40**
4.3.3. Китайская Народная Республика **40**
4.3.4. Республика Сингапур **42**
4.3.5. Соединенное Королевство Великобритании и Северной Ирландии **43**
4.3.6. Соединенные Штаты Америки **43**
4.3.7. Южная Корея **45**
4.3.8. Япония **45**
4.3.9. Австрия **45**

5. ВОЗМОЖНОСТИ СОВЕРШЕНСТВОВАНИЯ ИНСТРУМЕНТОВ РЕГУЛИРОВАНИЯ В РОССИИ 46

6. ОБЩИЕ ВЫВОДЫ И РЕКОМЕНДАЦИИ 48

Приложение **52**
Перечень публикаций для дополнительного изучения **52**

*21 марта 2022 года Тверской районный суд признал организацию Meta (социальные сети Instagram и Facebook) экстремистской, тем самым запретив ее деятельность в России

1. Актуальность проблематики и общее состояние исследований

В последние два-три года проблема дипфейков стала одной из актуальных тем для исследований в целом ряде развитых стран. Рост внимания к ней во многом обусловлен тем, что дипфейки по своей сути являются одним из результатов цифровой трансформации не только в сфере глобальных экономических процессов, но и в рамках меняющихся моделей социальных связей и общественного взаимодействия на самых различных уровнях. Развитию дипфейков, безусловно, способствует и процесс быстрого совершенствования программного обеспечения и аппаратных средств.

Вызовы, связанные с дипфейками, обсуждаются на таких площадках, как Организация Объединенных Наций (ООН), Международный союз электросвязи (МСЭ), «Группа двадцати», Организация по безопасности и сотрудничеству в Европе (ОБСЕ), Организация экономического сотрудничества и развития (ОЭСР), Институт инженеров электротехники и электроники, Международная организация по стандартизации (ИСО) и другие.

Дипфейки как воплощение новых технологических возможностей несут в себе серьезные риски. Спектр этих рисков весьма широк и продолжает увеличиваться. ДФ могут как представлять опасность для отдельного человека или группы людей, так и создавать более серьезные угрозы в информационном пространстве, затрагивая государственные интересы.

Происхождение технологии дипфейков связывают со студентом Стэнфордского университета Яном Гудфеллоу. В 2014 году, используя возможности машинного обучения (machine learning), он создал подобный продукт для разработчиков искусственного интеллекта[2].

Технология дипфейков быстро вышла за пределы узкого круга исследователей. В отчете Университетского колледжа Лондона[3], опубликованном в августе 2020 года, говорится, что фейковые аудио- и видеофайлы входят в топ-20 способов использования ИИ в преступных целях.

Также стоит отметить, что количество дипфейков в Интернете неуклонно растет, а единый подход к сбору статистических данных в этой области пока отсутствует. Так, к примеру, в октябре 2019 года на сайте телеканала CNN появилась статья[4], где говорилось, что в Сети тогда насчитывалось по меньшей мере 14 678 дипфейк-видео. При этом, согласно отчету компании Sensity «The State of Deepfakes»[5], количество обнаруженных ДФ-роликов в первые семь месяцев 2019 года составило 85 047. А к июню 2020-го, по данным Всемирного экономического форума, оно превысило 145 тысяч[6].

На семинаре ООН по новой геополитике ИИ[7], который прошел в 2018 году, было отмечено использование искусственного интеллекта в кампаниях по распространению поддельных новостей. В 2019 году исследователи инициативы ООН Global Pulse продемонстрировали, что речи ООН могут быть сфальсифицированы за 13 минут[8]. В 2020 году Всемирная организация интеллектуальной собственности заявила[9], что дипфейки способны вызвать серьезные проблемы, такие как нарушение прав человека, прав на неприкосновенность частной жизни, прав на защиту личных данных и так далее. А 25 августа 2021 года в Женеве в Институте ООН по исследованию проблем разоружения состоялось обсуждение, посвященное ДФ, доверию и международной безопасности[10].

[2] www.livescience.com/deepfake-ai.html
[3] www.crimesciencejournal.biomedcentral.com/counter/pdf/10.1186/s40163-020-00123-8.pdf
[4] www.edition.cnn.com/2019/10/07/tech/deepfake-videos-increase/index.html
[5] www.sensity.ai/reports
[6] www.weforum.org/agenda/2021/04/are-we-at-a-tipping-point-on-the-use-of-deepfakes
[7] www.un.org/en/academic-impact/un-hosts-seminar-new-geopolitics-artificial-intelligence
[8] https://genevasolutions.news/science-tech/deepfakes-are-getting-more-real-and-so-are-the-security-threats-un-dialogue
[9] https://wipo.int/export/sites/www/about-ip/en/artificial_intelligence/call_for_comments/pdf/org_aap.pdf
[10] https://unidir.org/events/2021-innovations-dialogue

За последние два года Управление перспективных исследовательских проектов Министерства обороны США (DARPA) потратило 68 миллионов долларов на разработки, призванные обнаруживать дипфейки[11].

Технология ДФ может иметь серьезные последствия как для национальной, так и для общественной безопасности[12]. Исследователи указывают на то, что она позволяет предоставлять ложную информацию в очень достоверной форме, тем самым манипулируя эмоциями людей и вызывая повсеместное недоверие.

Дипфейк представляет собой поддельный аудио- и/или визуальный контент, который создан с помощью генеративно-состязательных сетей (generative adversarial network, GAN)[13]. Одно из преимуществ GAN — способность совершенствоваться, используя набор обучающих данных, и создавать выборку с одинаковыми функциями и характеристиками. Например, GAN могут быть использованы для детального сканирования и сравнительного анализа реального и поддельного изображений или видео.

GAN состоят из двух взаимосвязанных компонентов. Первый, условный актер, пытается изучить статистические закономерности в наборе данных, таком как набор изображений или видео, а затем сгенерировать убедительные синтетические фрагменты информации. Второй, называемый критиком, пытается различить реальные и поддельные примеры. Обратная связь с критиком позволяет актеру создавать все более реалистичные модели.

При этом, согласно отчету международной антивирусной компании ESET о трендах информационной безопасности 2020 года[14], технологии ДФ продолжат активно развиваться.

[11] www.darpa.mil/program/media-forensics
[12] www. iopscience.iop.org/article/10.1088/1742-6596/1746/1/012068/pdf
[13] https://un.org/counterterrorism/sites/www.un.org.counterterrorism/files/malicious-use-of-ai-uncct-unicri-report-hd.pdf
[14] https://esetnod32.ru/company/press/center/eset-umnye-goroda-iskusstvennyy-intellekt-i-drugie-ib-trendy-2020-goda/

2. Общая классификация дипфейков

В рамках главы делается попытка систематизировать дипфейки по методам создания и создателям, по полезным и вредоносным, по целям и наносимому ущербу.

2.1. Использование термина

Термин **«дипфейк»** (deepfake, слияние понятий «глубокое обучение» и «подделка», «фальшивка») появился в конце 2017 года[15], произойдя от никнейма Deepfakes пользователя онлайн-платформы Reddit. Он применял инструменты машинного обучения с открытым исходным кодом, чтобы присоединять лица знаменитых женщин к телам женщин в порнографических видео и затем размещать такие поддельные клипы в соцсети[16]. Вскоре другой пользователь Reddit — Deepfakeapp — опубликовал приложение FakeApp, которое позволяло технически менее подкованным пользователям компьютеров создавать собственные дипфейки.

[15] www.internetjustsociety.org/legal-issues-of-deepfakes
[16] https://onlineethics.org/cases/deepfakes-and-value-neutrality-thesis

Термин **«дипфейк»** пока не имеет формального общепринятого технического определения[17]. В какой-то степени его можно отнести к разговорным словам. Однако он приобрел широкое распространение и прочно вошел в категорию общеупотребительной лексики.

В отчете Европейского парламента[18] дипфейки определяются как искусственные звуковые и/или визуальные медиа, кажущиеся подлинными. Их герои говорят или делают то, чего на самом деле никогда не было. Эти медиа созданы с использованием методов искусственного интеллекта, включающих машинное обучение (machine learning) и глубокое обучение (deep learning).

По своей природе **дипфейки** — это специальный алгоритм или набор алгоритмов, который анализирует множество параметров разнообразного контента, а затем, используя полученные данные, подменяет его в зависимости от поставленной изготовителем задачи. По сути, это процесс трансформации или синтеза, основанный на технологии искусственного интеллекта.

Корпорация Microsoft в 2022 году отметила, что распространение синтетических медиа станет еще изощреннее, дипфейки будут вполне убедительно чередоваться в Сети с реальными событиями, разворачивающимися в мире. В перспективе человечеству предстоит столкнуться с проблемой синтетических фальсификаций людей, поддельными беседами с ними по аудио- и визуальным каналам в режиме реального времени[19].

[17] www.theverge.com/2018/2/20/17032228/ai-artificial-intelligence-threat-report-malicious-uses
[18] www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf
[19] https://blogs.microsoft.com/on-the-issues/2022/05/03/artificial-intelligence-department-of-defense-cyber-missions/

2.2. Признаки и критерии, отличающие дипфейки от иных видов фейков

Создание сложных компьютерных изображений осуществлялось на протяжении десятилетий за счет использования различных визуальных эффектов в индустрии кино, телевидения и развлечений. Но последние достижения в области искусственного интеллекта привели к резкому увеличению реалистичности поддельного контента, а также к технической и финансовой доступности этих инструментов. Дипфейки следует отличать от так называемых дешевых подделок (cheapfakes)[20] и мелких подделок (shallowfakes)[21], а также от других фейков в Интернете, созданных без применения технологий искусственного интеллекта.

2.2.1. Дешевые (cheapfakes) и мелкие (shallowfakes) подделки

Помимо дипфейков, к формам манипуляции с аудио-, фото- и видеоконтентом относятся дешевые и мелкие подделки — cheapfakes и shallowfakes . Они создаются без применения искусственного интеллекта, с использованием недорогих компьютерных средств, порой даже при помощи базового программного обеспечения.

Примером может послужить резонансное видео с Facebook*, на котором спикер Палаты представителей США Нэнси Пелоси, выглядя пьяной, ругается во время выступления[22]. Дональд Трамп опубликовал этот клип в Twitter с подписью: «Пелоси заикается на пресс-конференции».

Видео было быстро признано подделкой. Только прежде его просмотрели миллионы человек, которые оставили почти 98 тысяч лайков. При этом Facebook* отказался[23] удалять ролик, решив отделаться лишь сокращением его распространения.

[20] www.datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1.pdf
[21] www.sciencemediahub.eu/2019/12/04/deepfakes-shallowfakes-and-speech-synthesis-tackling-audiovisual-manipulation/
* 21 марта 2022 года Тверской районный суд признал организацию Meta (социальные сети Instagram и Facebook) экстремистской, тем самым запретив ее деятельность в России
[22] https://www.theguardian.com/us-news/video/2019/may/24/real-v-fake-debunking-the-drunk-nancy-pelosi-footage-video
[23] https://www.reuters.com/article/us-facebook-deepfake-idUSKCN1TS023

Нижеприведенная схема иллюстрирует отличие дипфейков от дешевых подделок и содержит конкретные примеры.

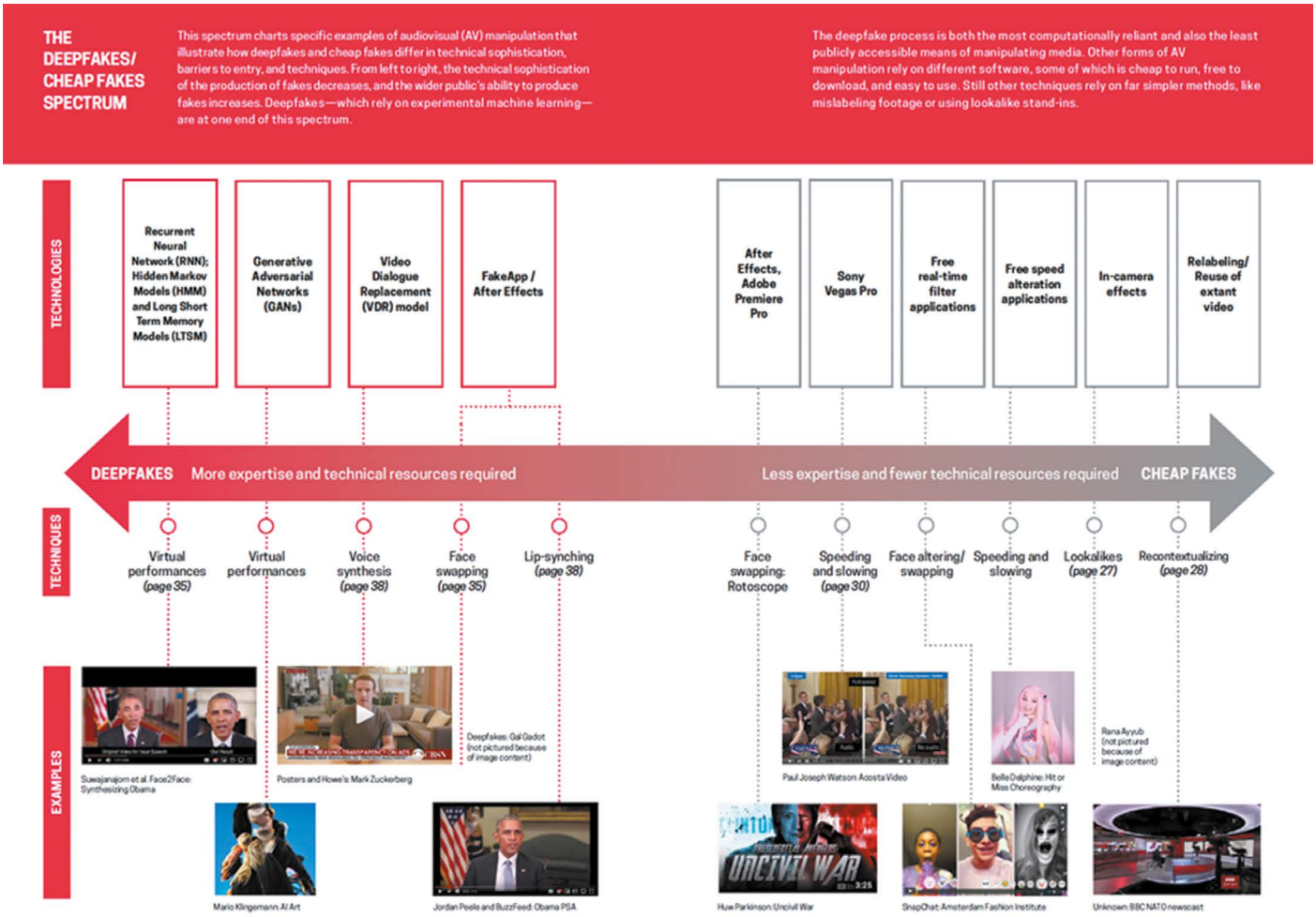


Рисунок № 1. Спектр дешевых подделок и дипфейков[24]

[24]Источник: https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1.pdf.

2.2.2. Типы фейков в Интернете

Для лучшего понимания такого явления, как дипфейк, приведем наиболее распространенную классификацию других различных фейковых материалов, которые могут быть созданы без технологий искусственного интеллекта.

Условно различают следующие виды таких фейков[25, 26]:

- Кликбейты в социальных сетях или на веб-сайтах в виде сообщений, статей, видео и другого вида информации. Основная цель — заставить как можно больше людей перейти по ссылке. Среди часто встречающихся приманок — фраза «вы не поверите, что...», привлекательные изображения, эмоциональный или юмористический тон, заявления о бесплатном и интригующем контенте.

- Фейковые твиты или вручную измененные ретвиты.

- Реклама, содержащая мошеннические или ложные утверждения.

- Создание условий для хакерской атаки, основные цели которой — кража личных данных, удержание «в заложниках» важного ресурса или секретной информации для получения выкупа и другие подобные правонарушения.

- Сенсационные заголовки, призванные вызвать интерес к содержанию материала, который на самом деле является рекламой или другого рода обманом.

- Вводящий в заблуждение контент — материалы с фальшивыми фактами. Применяется для формирования искаженного представления о персоне или явлении.

- Боты — фальшивые профили, в основном в социальных сетях. Создаются для распространения информации, и не всегда достоверной, при помощи автоматизированных технологий.

- Фишинг — ложный контент или поддельные двойники веб-сайтов, которые используются с целью получения личной информации.

- Поддельные аккаунты в социальных сетях для манипуляций общественным сознанием.

- Крупномасштабные мистификации[27] – преднамеренные недостоверные сообщения. Маскируются под новости и могут быть подхвачены и ошибочно подтверждены новостными агентствами.

- Дешевые и мелкие подделки[28] – синтезированный контент, полученный путем использования недорогих компьютерных средств либо базового программного обеспечения.

- Ложные аналитика и консалтинг[29].

- Фейковый спонсорский контент – имитация факта платного одобрения поста со стороны именитого бренда.

[25] www.internetmatters.org/wp-content/uploads/2020/11/Internet-Matters-Misinformation-Infographic-Final-1.pdf
[26] https://datajournalism.com/read/handbook/verification-1/additional-materials/types-of-online-fakes
[27] www.researchgate.net/publication/281818851_Deception_Detection_for_News_Three_Types_of_Fakes
[28] www.deepfakenow.com/what-is-the-difference-between-a-deepfake-and-shallowfake/
[29] www.youtube.com/watch?v=oaR9U-dA5Fc

2.2.3. Источники фейковых новостей

На основе анализа материалов, рассмотренных в рамках обзора, можно выделить пять основных категорий источников распространения фейковых новостей[30]:

- Поддельные веб-сайты. Главной задачей фальсификации популярных ресурсов, которым доверяют пользователи, чаще всего является использование личных данных людей в преступных целях.

- Веб-сайты, которые используют гиперболические или кликабельные заголовки и/или аналогичные описания своего контента в социальных сетях.

- Сатирические веб-сайты, которые предлагают критические комментарии политических и общественных явлений, но при этом распространяют их как реальные новости.

- Веб-сайты, которые распространяют вводящую в заблуждение и/или потенциально недостоверную информацию, предоставляют мнения в качестве новостей. Выглядят как традиционные электронные СМИ, но в действительности часто демонстрируют контент, подвергнутый манипуляциям.

- Компании, которые используют для недостоверного информирования службы обмена личными сообщениями, ботов, а также прибегают к атакам троллей в социальных сетях.

[30] www.qz.com/839160/heres-a-handy-cheat-sheet-of-false-and-misleading-news-sites/

2.3 Общая классификация дипфейков

Дипфейки можно условно классифицировать по методам создания, по создателям, на полезные и вредоносные, по типам злонамеренных и преступных целей и наносимому ущербу.

2.3.1. По методам создания

Классификация[31] дипфейков по методам их создания:

- смена лица;
- кукловодство;
- синхронизация губ;
- клонирование голоса;
- синтез изображений;
- генерация текста;
- передача нескольких стилей;
- подрисовывание;
- рендеринг.

Некоторые из перечисленных методов создания ДФ рассмотрены ниже.

Смена лица (face swaps) — это замена лица одного человека лицом другого либо аналогичная замена только его ключевых черт, а также манипуляция с фильтрами. Это общая особенность почти всех социальных сетей и приложений для видеочата.

Так, к примеру, приложение Snapchat позволяет изменять внешность с 2014 года. К примеру, используя технологию линз для распознавания лиц, оно дает возможность состариться.

Продукция всех подобных приложений квалифицируется как дипфейки. Ярким примером является ролик с основателем SpaceX Илоном Маском, в котором он якобы поет песню «Трава у дома»[32].

Кукловодство (puppeteering) — это видеоотрисовка определенных движений человека с помощью искусственного интеллекта путем создания 3D-моделей лица и тела.

В августе 2018 года Калифорнийский университет в Беркли представил доклад под названием Everybody Dance Now[33]. В нем исследовалась способность искусственного интеллекта применять движения профессионального танцора на видео тел танцоров-любителей.

В 2019 году японская компания Data Grid, занимающаяся технологиями искусственного интеллекта, создала продукт, который генерирует фотографии моделей, повторяющие изображения живых людей.

Синхронизация губ (lip-syncing) — это технология генерирования движений рта и мимики под заданное аудио. Цель таких алгоритмов искусственного интеллекта — имитировать речь определенного человека, в т. ч. и в тех условиях, в которых он никогда не находился.

Клонирование голоса (voice cloning) — это алгоритм создания синтетического голоса, похожего на исходный, для его последующего использования в генерировании речи определенного человека.

[31] https://towardsdatascience.com/ai-generated-synthetic-media-aka-deepfakes-7c021dea40e1?gi=4168371ae47b
[32] https://www.youtube.com/watch?v=iXtrGFrmIk4
[33] https://arxiv.org/pdf/1808.07371.pdf

Существует множество приложений и облачных сервисов для разработки синтетического голоса. Среди таковых — Microsoft Custom Voice, Lyrebird AI, iSpeech и VocaliD.

Синтез изображений — это метод генерирования новых визуальных элементов посредством использования технологий компьютерного зрения, глубокого обучения и GAN. Этот метод позволяет создавать любое компьютерное изображение, которое не является реальным.

Так, команда американской технологической компании NVIDIA создала сайт — генератор случайных лиц[34]. Для этого она использовала фото людей с фотохостинга Flickr.

Генерация текста — это метод автоматического создания любого вида текста посредством рекуррентных нейронных сетей (recurrent neural network[35]) или GAN.

Стоит отметить, что генерация текста развивает автоматизированную и роботизированную журналистику. А третье поколение алгоритма обработки естественного языка OpenAI и вовсе генерирует любой текст, включая табулатуры для гитары и компьютерный код.

Подрисовывание (inpainting) — система редактирования визуальных элементов, которая генерирует изображения с помощью масок произвольной формы, эскизов и цветов, предоставленных пользователями.

2.3.2. По создателям

Создателей дипфейков можно условно разделить как минимум на пять основных групп[36]:

- сообщества любителей;
- политические игроки, в том числе иностранные правительства и различные политактивисты;
- злоумышленники, в том числе мошенники;
- коммерческие организации;
- средства массовой информации.

Созданию большого количества любительских дипфейков способствует наличие базового программного обеспечения для сравнительно несложного процесса их изготовления. В целом, большая часть любителей воспринимают ДФ как новую форму онлайн-юмора и развлечений, а не как способ обмануть или запугать людей[37].

Однако дипфейки становятся и любимым инструментом многих хакеров. Количество незаконных и вредоносных фейковых видео, созданных на профессиональном уровне, удваивается примерно каждые шесть месяцев[38].

[34] www.thispersondoesnotexist.com/
[35] www.ibm.com/topics/recurrent-neural-networks
[36] www.timreview.ca/sites/default/files/article_PDF/TIMReview_November2019%20-%20D%20-%20Final.pdf
[37] https://edition.cnn.com/2019/06/22/tech/deepfake-for-fun/index.htm
[38] https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf

2.3.3. Полезные или безвредные дипфейки

Потенциально безвредные и полезные дипфейки можно найти в следующих областях[39]:

- **Коммерческое использование, аудиографическая продукция и индустрия развлечений.** Например, сходства синтетических голосов и лиц с реальными можно использовать в фильмах для достижения творческого или юмористического эффекта. Также благодаря им есть шанс сохранять связность историй, если сами артисты недоступны. Речь идет о тех ситуациях, когда актер в связи с непредвиденной ситуацией не может лично явиться на съемочную площадку, а приостанавливать работу над фильмом невозможно или когда режиссеры воспроизводят голос умершего артиста. В подобных ситуациях необходимо соблюдать авторское право и предпринимать меры для защиты персональных данных.

- **Человеко-машинное взаимодействие.** Приобретение опыта в технологиях, где люди трудятся в команде с машинами.

- **Видеоконференции.** Например, британская аудиторско-консалтинговая компания Ernst & Young использует технологию дипфейков для видеообщения «сотрудников» с клиентами вместо личных встреч[40].

- **Личное и творческое выражение.** Дипфейки используются для создания в сети «Интернет» аватаров. Благодаря им многие люди расширяют свои возможности, реализуют свои идеи, держатся за свои убеждения и получают простор для самовыражения. Например, люди с физическими недостатками могут использовать синтетические аватары для более позитивной активности в Интернете.

- **Свобода слова.** Дипфейки позволяют правозащитникам и журналистам оставаться анонимными в условиях диктаторских режимов. В некоторых ситуациях ДФ используются для защиты конфиденциальности лиц, дающих острые комментарии.

- **Здравоохранение.** Дипфейки дают возможность разрабатывать новые способы лечения и диагностики заболеваний. К примеру, специалисты американской технологической компании NVIDIA в соавторстве с крупнейшими медицинскими исследовательскими центрами Clinical Data Science и Mayo Clinic посредством алгоритмов искусственного интеллекта создали синтетическое изображение головного мозга с опухолями, а затем, используя его, научили ИИ обнаруживать этот вид заболеваний с достоверностью, которая присуща алгоритмам, обученным на реальных изображениях. Кроме того, дипфейки способны воссоздавать в цифровом виде ампутированные конечности или помогать трансгендерам воочию видеть себя в предпочитаемом поле.

[39] https://scholarship.law.bu.edu/cgi/viewcontent.cgi?article=1640&context=faculty_scholarship, https://towardsdatascience.com/positive-use-cases-of-deepfakes-49f510056387
[40] https://www.respeecher.com/blog/deepfakes-workplace-synthetic-media-improves-life-office

- **Приватность.** Пользователи могут менять до неузнаваемости свое лицо в разного рода чатах, чтобы их реальная внешность оставалась в секрете в целях личной безопасности.

- **Образование.** Дипфейки создают множество полезных возможностей для преподавателей. Учебная информация может демонстрироваться более наглядно по сравнению с традиционными средствами, такими как стандартные семинары и лекции. В этом смысле дипфейки в чем-то схожи с уже пройденной волной образовательных инноваций, когда в учебном процессе стали широко использоваться видеоматериалы и другие средства визуализации учебных курсов. С помощью ДФ можно создавать ролики с историческими фигурами, говорящими непосредственно со студентами. Например, проект Dimensions in Testimony Фонда Шоа Университета Южной Калифорнии[41] привлек большое внимание средств массовой информации, поскольку в нем были представлены интервью и голографические записи 15 выживших во время холокоста.

- **Электронная коммерция.** Дипфейки позволяют покупателям использовать собственные образы для виртуальной примерки одежды, аксессуаров и не только. Для этого лишь нужно подставить к цифровым моделям свое лицо и выбрать подходящий тип телосложения. Яркий пример подобных возможностей — приложение Superpersonal[42]. Благодаря ему обычные гости Недели моды в Лондоне смогли, участвуя в виртуальных показах брендов, почувствовать себя топ-моделями[43].

- **Коммуникация.** Благодаря дипфейкам можно создать впечатление, будто человек говорит на иностранном языке, как на родном. Таким образом, например, лидеры мнений способны в критических ситуациях с куда большей вероятностью достучаться до сердец поклонников из разных стран. Так к пользе дипфейков прибегал английский футболист Дэвид Бекхэм в сотрудничестве с британской благотворительной организацией здравоохранения, когда записывал ролик, призывающий бороться с малярией. Благодаря дипфейк-технологиям Бекхэм без особых лингвистических знаний смог обратиться к поклонникам из девяти стран на их родных языках[44].

[41] https://sfi.usc.edu/dit
[42] www.producthunt.com/posts/superpersonal
[43] https://www.fialondon.com/projects/hanger-x-superpersonal/
[44] https://www.techtarget.com/whatis/definition/deepfake

2.3.4. Вредоносные дипфейки

Прогресс в технологии дипфейков открывает множество возможностей для самых разных целей — как с положительными, так и с отрицательными последствиями. Большая часть вреда, причиняемого ДФ, подпадает под различные правонарушения, предусмотренные законодательствами стран. Однако оставшаяся часть потенциального вреда пока не охвачена законными актами, но может быть покрыта внесением поправок в соответствующие правовые инструменты.

Технология дипфейков при своей технической нейтральности может нести вредоносный и даже разрушительный потенциал на индивидуальном и коллективном уровнях, на уровне организаций и даже в масштабе страны. В совокупности ДФ способны привести к политической и социальной поляризациям, общественной нестабильности, вражде, недоверию, нагнетанию паники или истерии, агрессии, насилию и не только.

По сути, при злонамеренном использовании дипфейки могут быть инструментом создания практически всех видов деструктивного контента в Сети и информационной среде в целом. Общая классификация основных видов деструктивного контента представлена в «Сравнительно-правовом анализе мер по противодействию распространению противоправного (деструктивного) контента в сети Интернет», подготовленном CGITC в августе 2021 года[45].

Наиболее характерные негативные и вредоносные дипфейки и их проявления будут рассмотрены ниже.

2.3.4.1. По типам злонамеренных и преступных целей

Ввиду быстрого распространения дипфейки могут использоваться для целого ряда злонамеренных и преступных целей, среди которых[46]:

- нарушение прав человека;
- подрыв репутации;
- домогательство;
- изготовление порнографии;
- шантаж;
- мошенничество;
- нарушение авторских прав;
- разглашение конфиденциальной информации;
- вторжение в частную жизнь;
- подделка документов (например, морфинг-атака (morphing attack)[47], или манипуляция с изменением лица (face-morphing trend), позволяет нескольким людям использовать один паспорт посредством процесса, который объединяет их фотографии в одну, но при этом сохраняет сходство со всеми «владельцами» документа);
- умышленное причинение душевных страданий;
- нарушение прав потребителей;
- фальсификация онлайн-идентичности и обман механизмов KYC[48];
- манипулирование электронными доказательствами при расследовании уголовных дел;
- разжигание социальных волнений и политической поляризации;
- фальсификация геопространственных данных в военных конфликтах[49].

[45] www.cgitc.ru/groups/analiticheskie-issledovaniya/kratkiy-sravnitelno-pravovoy-analiz-mer-po-protivodeystviyu-rasprostraneniya-protivopravnogo-destruk/
[46] https://unicri.it/sites/default/files/2020-11/Abuse_ai.pdf
[47] https://www.europol.europa.eu/publications-documents/malicious-uses-and-abuses-of-artificial-intelligence
[48] KYC — принцип работы финансовых институтов, который обязывает их идентифицировать личность человека перед тем, как тот сможет проводить операции
[49] www.tandfonline.com/doi/abs/10.1080/15230406.2021.1910075

2.3.4.2. По наносимому ущербу

При рассмотрении негативных последствий в первую очередь стоит выделить следующие группы дипфейков:

- наносящие психологический ущерб;
- причиняющие финансовый ущерб;
- наносящие социальный ущерб.

Таблица. Классификация дипфейков по типу наносимого ущерба

Психологический ущерб	Финансовый ущерб	Социальный ущерб
Сексторция	Вымогательство	Манипуляции со СМИ
Клевета	Кража личных данных	Ущерб экономической стабильности
Запугивание	Мошенничество	Ущерб системе правосудия
Издевательства	Манипулирование ценами на акции	Ущерб научной системе
Подрыв доверия	Ущерб бренду	Подрыв доверия
	Репутационный ущерб	Ущерб демократии
		Манипулирование выборами
		Ущерб международным отношениям
		Ущерб национальной безопасности

Источник: «Борьба с дипфейками в европейской политике», Научно-исследовательская служба Европейского парламента[50]

[50] https://www2.deloitte.com/us/en/pages/advisory/articles/risk-in-the-digital-era.html

3. Вызовы и наиболее проблемные вопросы, связанные с дипфейками

В сегодняшнем массиве информации трудно провести границу между реальным и фальшивым. Поддельные тексты, изображения, аудио- и видеоматериалы усугубляют и без того сложное ориентирование человека в информационной среде, подрывают доверие, создают в обществе ситуацию, когда трудно найти факты, на которые можно опереться при анализе действительности и формировании картины мира.

В исследовании международной сети консалтинговых и аудиторских компаний Deloitte о рисках в цифровую эпоху[51] отмечается, что широкомасштабное распространение поддельной и непроверенной информации посредством онлайн-платформ и передовых инструментов манипулирования, включая дипфейки, ведет к новым типам информационных войн.

Также стоит отметить, что часть инструментария для производства дипфейков доступна всем, достаточна проста в использовании и поэтому может быть освоена пользователями без специальных технических знаний.

[51] https://www2.deloitte.com/us/en/pages/advisory/articles/risk-in-the-digital-era.html

3.1. Угрозы

В последующих подпунктах приведены основные угрозы дипфейков для государства, бизнеса и общества.

3.1.1. Политика

В связи с широкой вепонизацией дипфейков[52] их распространение несет серьезные угрозы для национальной безопасности. Предоставляя общественности ложную информацию в достоверной форме и тем самым манипулируя эмоциями людей, ДФ вызывают повсеместное недоверие в обществе, а также вмешиваются в выборы.

Кроме того, дипфейки оказывают разрушительное влияние на геополитику и межгосударственные отношения. Сегодня на фоне многочисленных подделок в информационной среде страны выдвигают взаимные обвинения и порой ведут напряженный диалог, оперируя якобы обличительными фактами и неопровержимыми доказательствами.

За счет подрыва доверия людей к потенциальным и действующим политическим фигурам или государственным институтам дипфейки разжигают ненависть и способствуют росту терроризма[53].

[52] www.realinstitutoelcano.org/en/analyses/the-weaponisation-of-synthetic-media-what-threat-does-this-pose-to-national-security/
[53] https://link.springer.com/article/10.1007/s44206-022-00010-6

3.1.2. Судебная система

ДФ способны подорвать доверие общества и к государственным институтам, включая систему правосудия. Так, в материале «Дипфейки: угроза конфиденциальности, демократии и национальной безопасности» электронного репозитория научных статей и препринтов SSRN[54] фальсификация доказательств в судебной системе определяется одной из основных угроз, исходящих от ДФ. Ложная информация, представленная в качестве аргумента со стороны защиты или обвинения, может серьезно повлиять на ход разбирательства.

Поэтому вопрос аутентификации цифровых доказательств в судебной практике считается одним из ключевых на повестке дня. Чаще всего проблемы возникают во время перекрестных допросов: одна сторона дает утвердительные показания относительно деталей дипфейка, а другая — отрицает его подлинность. Это возлагает большую, чем обычно, нагрузку на судебных экспертов и требует дополнительных финансовых расходов на проверку фактов.

Например, в британском деле[55] об опеке над детьми в качестве доказательства в суде мать представила дипфейковый аудиофайл. Посредством специальной программы из сети «Интернет» она воссоздала голос их отца и имитировала поток угроз в адрес детей, чтобы подтвердить его жестокость. Однако после экспертизы фальшивый аудиоматериал не был принят к рассмотрению.

Кроме того, сегодня многие суды признали онлайн-формат рассмотрения дел. И это влечет дополнительные вызовы[56]. Посредством дипфейк-технологий истцы, ответчики или свидетели таким образом во время видеосвязи могут быть вовсе не теми, за кого себя выдают.

3.1.3. Социальная инженерия

Большая часть источников дает объяснение термину «социальная инженерия» в контексте следующих областей:

- информационная безопасность (психологическое манипулирование с целью доступа к информации);
- социология (целенаправленное формирование общественного поведения).

Применительно к проблеме ДФ понятие «социальная инженерия» вероятно следует рассматривать несколько шире. По сути, это создание возможностей для манипулирования большими общественными группами посредством широкого применения как технических, так и психологических механизмов. Здесь дипфейки выступают в качестве эффективного инструмента воздействия на социальное поведение людей, инструмента внушения и внедрения ожидаемых установок и реакций. По мнению экспертов, последние успехи в области технологий выводят исторические методы социальной инженерии на новый уровень как по масштабу, так и по сложности[57].

После массированной экспансии в сферу развлечений и досуга ДФ начали все активнее проникать в область политики и общественных отношений. Ложные сообщения, кампании по дискредитации лидеров мнений и знаменитостей, целенаправленная манипуляция аудиторией, призывы населения к тем или иным действиям — сегодня все это не обходится без дипфейков.

[54] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954
[55] www.abajournal.com/web/article/courts-and-lawyers-struggle-with-growing-prevalence-of-deepfakes
[56] https://static1.squarespace.com/static/5c1bfc7eee175995a4ceb638/t/5f1b23e97ab8874a35236b67/1595614187464/Final+white+paper+pdf.pdf
[57] https ://mysecuritymarketplace.com/reports/social-engineering-blurring-reality-and-fake-a-guide-for-the-insurance-professional/

При этом качество технического исполнения поддельных материалов улучшается: они становятся все более сложными для выявления.

Нельзя обойти стороной и такое новое явление, как компрометация деловой идентификации (business identity compromise)[58]. В этой связи предполагается создание синтетических корпоративных личностей или эмуляция существующего сотрудника.

Консалтинговая и аудиторская международная сеть компаний PwC провела исследование, в результате которого ввела термин «удаленная онлайновая социальная инженерия» (remote online social engineering)[59]. В современном мире существует множество технических приемов, подходящих под это определение. Значительная часть из них де факто используется киберпреступниками, спектр целей которых сегодня тоже значительно увеличился. Например, помимо несанкционированного доступа к данным и кражи конфиденциальной информации, злоумышленники переманивают квалифицированных специалистов, ищут слабые места в управлении, создают конфликты в коллективе.

Важно еще раз подчеркнуть, что социальная инженерия в качестве слабого звена рассматривает именно человеческий фактор, поэтому технические приемы используются в сочетании с психологическими в отношении как конкретного индивида, так и группы людей. Принимая во внимание данное обстоятельство, можно констатировать, что для защиты от дипфейков должны использоваться не только технический арсенал и набор правил безопасности, но и психологическая подготовка. Это относится как к персоналу компаний, так и к самому широкому кругу пользователей Сети, которые могут стать жертвами обмана и манипуляции. В данном случае важными становятся такие качества, как готовность к критическому мышлению, способность взвешенно и трезво реагировать на нештатные ситуации и сценарии, требующие быстрых решений.

3.1.4. Финансовая сфера

В зрелых и здоровых экономиках технологически развитых стран дипфейки не представляют серьезной угрозы для стабильности финансовой системы в целом, но они могут причинить вред отдельным людям, незащищенным предприятиям и организациям, в том числе и государственным регулирующим органам[60].

Например, использование видеосвязи для удаленного открытия счетов банковским клиентам в связи с дипфейковыми возможностями определенно сопряжено со снижением безопасности. Такого мнения придерживается Банк России, который назвал ДФ угрозой для идентификации по видеосвязи[61].

3.1.5. Бизнес

Дипфейки способны разрушить положительную репутацию благонадежного предприятия или стать пособниками мошенников в обмане фирм на крупные суммы денег[62]. Также они могут использоваться для манипулирования рынком и акциями, вызывать панику на биржах, влиять на волатильность валют и не только. Например, фальшивое видео может продемонстрировать широкой публике, как глава некой корпорации опроверг важное слияние или заявил о банкротстве[63].

[58] https://blog.knowbe4.com/fbi-warns-against-deepfakes-potential-for-social-engineering
[59] https://www.pwc.co.uk/issues/cyber-security-services/insights/dark-art-of-remote-online-social-engineering.html
[60] https://carnegieendowment.org/files/Bateman_FinCyber_Deepfakes_final.pdf
[61] https://www.forbes.ru/newsroom/tehnologii/429961-cb-nazval-dipfeyki-ugrozoy-pri-identifikacii-klientov-po-video
[62] https://www.theverge.com/2019/9/5/20851248/deepfakes-ai-fake-audio-phone-calls-thieves-trick-companies-stealing-money
[63] www.timreview.ca/sites/default/files/article_PDF/TIMReview_November2019%20-%20D%20-%20Final.pdf

3.1.6. Цифровые платформы и социальные сети

Социальные сети и цифровые платформы являются тем пространством, где наиболее активно и массово распространяются дипфейки. Безусловно это вызывает озабоченность и у самих владельцев этих ресурсов из-за увеличивающегося количества жалоб, рекламаций и судебных исков. В этих условиях крупные технологические компании (платформы и экосистемы) собственными силами и в сотрудничестве с другими заинтересованными игроками создают различные инструменты и решения для автоматизированного обнаружения ДФ, а также принимают соответствующие политики[64], препятствующие распространению (масштабированию в сети) вредоносных ДФ.

3.1.7. Дети

Дипфейки способствуют росту и усугублению способов издевательств и онлайн-насилия над детьми в Интернете. Также хоть и ложное, но реалистичное изображение сцен агрессии, жестокости или откровенных половых актов может нанести ущерб психическому здоровью ребенка.

Одно из самых популярных занятий современных детей — создавать видеоролики о себе и размещать их в различных социальных сетях, а также генерировать свою внешность в зрелом возрасте. Загружая свои фотографии на подобные платформы, несовершеннолетние неосознанно предоставляют их третьим лицам, способствуют созданию обширных библиотек контента, которые впоследствии часто используются для производства дипфейков, в том числе сексуального характера.

Самой большой проблемой в расследовании инцидентов с контентом, на котором визуализировано сексуальное насилие над детьми, считается сложность с идентификацией ребенка, фигурирующего в нем[65]. Дипфейк-технологии позволяют скрывать лица, а не установив жертву, весьма сложно выйти на преступника. Также существует риск того, что следователи зря потратят время на поиск несовершеннолетнего, который, как окажется в итоге, на самом деле не подвергался сексуальному насилию, а лишь стал участником сфальсифицированного ролика.

Сегодня приложения для создания дипфейков доступны для скачивания на любом смартфоне. Поэтому сейчас остро стоит вопрос о необходимости введения строгих обязательств по их использованию, чтобы они могли применяться только в законных целях. Многие эксперты считают, что именно разработчики подобных приложений, например, таких как Reface и FaceMagic, должны нести ответственность за ненадлежащее применение их продукции.

3.1.8. Доверие к информации

Дипфейки могут производить эффект, который называют дивидендами лжеца (the liar's dividends). Его суть заключаются в том, что люди отрицают подлинность действительно правдивого контента, так как во всем пытаются разглядеть ДФ-технологии.

[64] www.ntrepidcorp.com/managed-attribution/deepfakes-social-media/
[65] www.respect.international/wp-content/uploads/2019/02/NetClean_Report_2018.pdf

Так, к примеру, масштабное распространение синтетических средств массовой информации побуждает общество называть реальные СМИ фейковыми. При этом эксперты Фонда имени Конрада Аденауэра считают, что «дивиденды лжеца» будут прогрессировать и дальше[66].

По мнению[67] Кай-Фу Ли, доктора наук, венчурного инвестора и автора книг «ИИ-2041» и «Сверхдержавы искусственного интеллекта», мир столкнется с риском неотличимости правды от лжи, как для отдельного человека, так и для социумов любых масштабов. Достигнутый к началу 2023 г. уровень развития инструментов ИИ-генерации контента (AI Content Generator Tools) таков, что начнется их внедрение в поисковых системах и онлайн торговле товарами и услугами. Онлайн системы нового типа будут обладать возможностью убедительного обоснования навязываемого человеку выбора, с учетом его персональных предпочтений, вкусов, склонностей, пристрастий и предубеждений. Преимущества такого выбора будут представлены в наиболее привлекательном для человека сочетании мультимедийных форматов.

3.1.9. Дезинформация

Дипфейки часто дезинформируют людей. В связи с этим на международных площадках растет озабоченность по этому поводу.

ОЭСР в документе 2022 года «Распутывание неправды в Интернете: создатели, распространители и как их остановить»[68] под дезинформацией предлагает понимать достоверно ложную или вводящую в заблуждение информацию, которая сознательно создается и распространяется для получения экономической выгоды, манипулирования или причинения вреда лицу, социальной группе, организации или стране. При этом ОЭСР отмечает, что фейковые новости, синтетические медиа, включая дипфейки, и мистификации среди прочего являются формами дезинформации.

Дезинформация представляет собой угрозу целостности информационной экосистемы. По данным ЮНИСЕФ за 2021 год, в социальных сетях ложная информация распространяется быстрее, глубже и шире, чем правдивая, и часто входит в число самых популярных сообщений[69].

Стоит отметить, что угрозы дезинформации в Интернете обсуждались на прошедшем в декабре 2021 года круглом столе ОБСЕ «Дипфейковые новости: искусственный интеллект и дезинформация как вызов многосторонней политике»[70]. На мероприятии с участием большого количества международных экспертов были разобраны различные аспекты борьбы с дезинформацией посредством модерации контента, а также высказаны мнения по целому ряду вызовов, связанных с распространением ложной информации на онлайн-платформах и в глобальной цифровой экосистеме.

Важно понимать, что сочетание новых форм дезинформации с вирусным распространением контента в Интернете является источником беспрецедентных проблем. Так, использование дипфейков может усилить кампании по дезинформации физических лиц с целью оказания воздействия на избирателей[71].

[66] www.kas.de/documents/252038/7995358/Deepfakes+-+Eine+Bedrohung+f%C3%BCr+Demokratie+und+Gesellschaft.pdf/c4c7bc69-a5b6-8141-dca1-bb1f6869f806
[67] www.youtube.com/watch?v=JGiLz_Jx9uI&t=11510s
https://dzen.ru/a/Y7RwQ1AsKBkGwmIi
[68] www.oecd-ilibrary.org/docserver/84b62df1-en.pdf
[69] https://www.unicef.org/globalinsight/media/2096/file/UNICEF-Global-Insight-Digital-Mis-Disinformation-and-Children-2021.pdf
[70] www.osce.org/representative-on-freedom-of-media/506347

Например, можно сфабриковать сюжет, где политический противник берет взятку. Таким образом, политическая дезинформация угрожает способности электората достоверно оценивать государственных чиновников и выбирать компетентных лидеров.

Кроме того, использование дипфейков для дезинформации несет в себе угрозы для демократических механизмов[72], снижает доверие к институтам и власти[73]. Поэтому в связи со всем вышесказанным регуляторы стоят перед необходимостью не только создавать барьеры для распространения вредоносных дипфейков, но и внедрять в практику различные поощрения ответственного поведения акторов при производстве контента и его передаче конечным пользователям.

3.2. Отдельные технические вопросы, связанные с дипфейками, некоторые аспекты обнаружения ДФ

Комплексная система защиты от деструктивного использования дипфейков прежде всего зависит от применения современного программного обеспечения, способного в том числе различать настоящий и фейковый контент. Технические меры для борьбы с вредоносными ДФ включают создание специальных систем искусственного интеллекта и развитие криптографических методов, которые могут быть интегрированы в видео- и аудиозаписывающее оборудование с целью выявления несанкционированного доступа.

Так, процесс обнаружения визуального дипфейка обычно включает три этапа:

- проверку файла на наличие признаков объединения двух или более изображений или видео;
- анализ характеристик освещенности и других физических свойств контента на проявления подделки;
- отслеживание логических несоответствий, таких как погода, несоответствующая дате, или фон, не коррелирующий с местоположением.

Если затрагивать вопрос обнаружения дипфейков более подробно, то можно выделить возможность искусственного интеллекта распознавать частоту сердечных сокращений по так называемому обесцвечиванию лица человека на видео или изображении[74]. Когда в разных частях лицах частота не совпадает, специалист делает вывод, что перед ним ложный контент.

[71] www.un.org/ru/85483
[72] https://www.un.org/counterterrorism/sites/www.un.org.counterterrorism/files/countering-terrorism-online-with-ai-uncct-unicri-report-web.pdf
[73] www.allianceofdemocracies.org/wp-content/uploads/2020/04/Disinformation-Deepfakes-Democracy-Waldemarsson-2020.pdf
[74] www.thedeepfake.report/en/en-what-are-deepfakes

Также в этой связи стоит обратить внимание на технологию блокчейн[75]. Используя цепочки блоков данных в децентрализованной сети и сопоставляя информацию на выходе с определенным стартовым неизменным ключом, она может отследить путь цифрового контента вплоть до его создателя.

Многие эксперты даже предлагают сформировать на государственных уровнях органы, которые будут отслеживать дипфейки посредством технологии блокчейн[76].

Еще один пример решения для борьбы с дипфейками — технология Amber Authenticate. Работая на устройствах в фоновом режиме, она запоминает оригинальные данные контента, пока человек взаимодействует с камерой или диктофоном, а затем архивирует эти цифровые отпечатки для общего доступа[77].

При этом исследователи отмечают, что при использовании всевозможных технологий сегодня точность обнаружения, например, дипфейковых изображений все-таки пока не на уровне идеала. Предположительно, она составляет 99,3% для необработанного контента и 81% для более сложных задач с низким качеством[78].

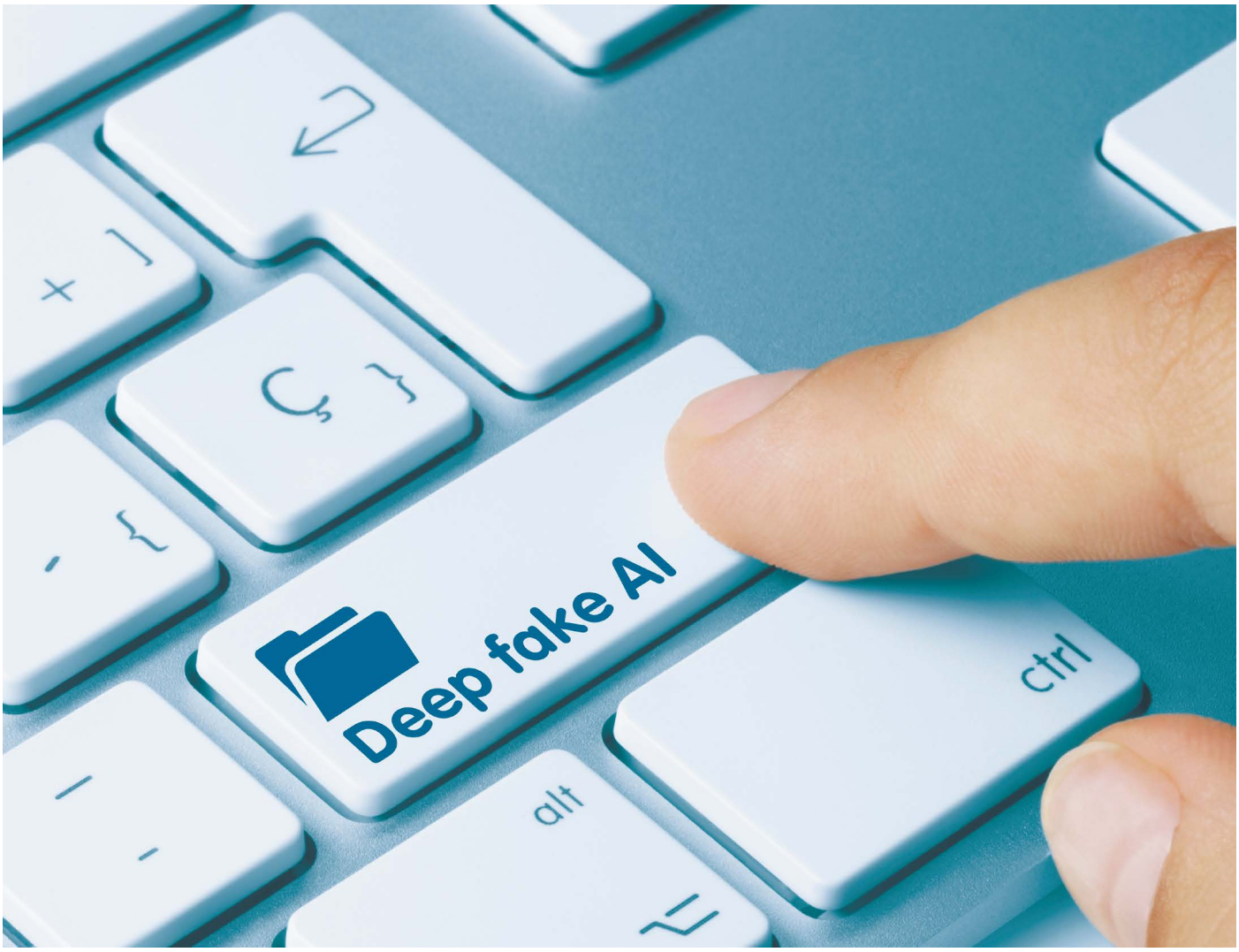
Кроме того, специалисты выделяют еще одну проблему. Дело в том, что факт выявления противоправного дипфейка не всегда ведет к идентификации его создателей, а соответственно, и к привлечению их к ответственности.

Нельзя не отметить, что даже самые современные средства выявления дипфейков и противодействия им всегда будут вынуждены догонять вновь возникающие и все более изощренные способы внедрения в массы ложного контента. Поэтому ряд международных исследователей полагают, что наилучшим результатом в вопросе противодействия ДФ может стать разработка именно правовых инструментов для управления рисками.

Однако, по мнению экспертов, ответственность за причиняемый вред должны нести именно отдельные лица — всеобщие запреты на новую, часто полезную, технологию вводить не следует. Отмечается также, что пока государственные и регулирующие органы не предпринимают достаточных усилий для обеспечения подлинности контента в публичном пространстве. Побуждение цифровых платформ к самоцензуре приносит лишь ограниченный эффект.

[75] www.researchgate.net/publication/331769696_Combating_Deepfake_Videos_Using_Blockchain_and_Smart_Contracts
[76] www.wired.com/story/the-blockchain-solution-to-our-deepfake-problems/
[77] www.natlawreview.com/article/rise-deepfake-demands-urgent-legal-reform-uk
[78] www.arxiv.org/abs/1901.08971

4. Наиболее значимые зарубежные практики и подходы



4.1. Исследования и инициативы

Правовая база стран недостаточно подготовлена для реагирования на технологические вызовы, связанные с развитием ДФ-технологий. Поэтому фактор быстрого распространения цифрового контента через различные онлайн-платформы сейчас приводит к возникновению разработок для обнаружения ложного контента по инициативе бизнеса, академических и политических кругов.

В мире наблюдается рост материалов, посвященных различным аспектам разработки и использования дипфейков. Такие материалы можно разделить на научно-прикладные, которые в основном посвящены инженерным вопросам, научно-популярные и публицистические статьи, а также всевозможные обзоры по существующим и новым видам ДФ, в том числе содержащие анализ отдельных юридических аспектов проблематики.

Среди актуальных исследований можно выделить материалы следующих организаций:

- **Белферский центр науки и международных отношений[79] Гарвардской школы Кеннеди.**
  В его брошюре «Дипфейки для политиков и регуляторов» 2020 года[80] ДФ определяются как синтетические аудиальные или визуальные носители, разработанные с использованием методов глубокого машинного обучения, которые настолько реалистичны, что способны обмануть аудиторию. При этом они крайне разнообразны по технической сложности и применению. Шкала начинается с низкокачественных дешевых дипфейков, а заканчивается высококачественной продукцией, способной оказать воздействие на восприятие реальности человеком, повлиять на его сознание и поведение.

Как отмечают исследователи Белферского центра науки и международных отношений, разработка синтетического аудиовизуального контента не нова. Голливудские кинематографисты используют компьютерные изображения (computer-generated imagery) с 1970-х годов с целью усиления правдоподобности происходящего на экране и снижения степени критического восприятия зрителей к вымышленным образам. А развитие цифровых технологий сделало сложные синтетические носители недорогими и простыми в производстве, особенно благодаря распространению бесплатного программного обеспечения с открытым исходным кодом.

- **Фонд информационных технологий и инноваций США[81].**
  Вице-президент Фонда Даниэль Кастро в своей статье 2020 года для журнала Government Technology[82] анализирует подходы законодателей штата Нью-Йорк к будущему регулированию дипфейк-технологий. По мнению автора, одна из главных проблем заключается в том, что ДФ используются в рамках кампаний дезинформации с очевидной целью повлиять на результаты выборов.

Другая серьезная проблема, по мнению создателя материала, — использование ИИ для производства порнографических изображений или видео с участием знаменитостей, в основном женщин, без их согласия.

Как указывает Кастро, в исследовании, проведенном в 2019 году нидерландской компанией Deeptrace[83], которая борется с фейками в Интернете, говорится, что 96 % обнаруженных ДФ содержали порнографический контент без очевидного согласия людей на использование их образов. При этом к моменту публикации результатов исследования в совокупности синтетические видео сексуального характера на четырех популярных специализированных сайтах просмотрели около 134 миллионов человек.

Еще один немаловажный вопрос, с которым сталкиваются законодатели, — это способы защиты прав в сфере контроля за коммерческим использованием имиджа человека в цифровой среде. Особенно остро он встает, когда речь заходит о знаменитостях, привыкших взимать плату за любое применение их образа.

[79] https://www.belfercenter.org/
[80] https://www.belfercenter.org/sites/default/files/2020-10/tappfactsheets/Deepfakes.pdf
[81] www.itif.org/
[82] www.govtech.com/policy/deepfakes-are-on-the-rise-how-should-government-respond.html
[83] https://deeptracelabs.com/mapping-the-deepfake-landscape

По этой причине члены законодательного собрания штата Нью-Йорк рассматривали, но в итоге все-таки не приняли закон, ранее поддержанный Гильдией киноактеров США. Проект закона предполагал введение законодательных правил на публичность для отдельных категорий лиц. В частности, обсуждалось положение, которое запрещало бы использовать образы умершего человека без его заранее оформленного согласия или в ином случае без одобрения его наследников.

- **Научный журнал Illinois Law Review[84].** В юридическом обзоре «Дипфейки. Гусь приготовлен?» («AI Deepfakes. The Goose Is Cooked?»)[85] 2020 года издание отметило, что дипфейки представляют серьезную угрозу в суде в моменты определения подлинности важнейших доказательств, поскольку непрерывное развитие технологий значительно усложняет процесс верификации цифровых аргументов. В судебной практике были случаи, когда обвиняемые в хранении детской порнографии утверждали, что найденные у них записи были синтетическими, то есть на самом деле в реальный сексуальный процесс изображенные на них дети не вовлекались.

- **Независимый глобальный аналитический центр Observer Research Foundation[86].** В статье «Обсуждение этики дипфейков»[87] 2020 года говорится о том, что ДФ ускоряют процесс падения доверия населения к СМИ, что способствует расцвету релятивизма, разрушает структуру демократии и гражданского общества. С другой стороны, наличие в цифровом пространстве синтетических медиа позволяет лидерам мнений списывать на фальсификацию любую неудобную им правду. По мнению авторов статьи, вкладывание определенных слов в чужие уста, замена чьего-либо лица другим, создание ложных изображений и цифровых клонов известных личностей с целью введения людей в заблуждение являются этически сомнительными действиями, которые должны быть наказуемы.

Нравственные аспекты производства и распространения дипфейков становятся еще более запутанными, когда речь заходит о синтетической порнографии, произведенной по обоюдному согласию актеров — цифровых марионеток, созданных по образам реальных людей.

Такие согласованные и поэтому уже фактически легализованные порнографические синтетические медиа могут получить распространение, что, безусловно, будет иметь негативный эффект.

Еще одна область, вызывающая беспокойство, — это синтетическое воскрешение человека. Главной задачей, требующей немедленного решения, является определение принадлежности прав на образ и голос людей после их смерти. Существуют опасения этического характера по поводу того, что дипфейки могут быть использованы, например, с целью подрыва репутации известных политических деятелей после их кончины для достижения их оппонентами определенных целей.

[84] www.illinoislawreview.org/
[85] www.illinoislawreview.org/blog/ai-deepfakes/
[86] www.orfonline.org
[87] www.orfonline.org/expert-speak/debating-the-ethics-of-deepfakes/

Кроме этого, сегодня уже существуют компании, которые создают синтетические голоса родственников в качестве нового вида терапии после тяжелой утраты близких. Сторонники синтетического воскрешения заявляют, что это аналогично хранению фотографий или видео умершего, однако надо понимать, что не все люди обладают должной ответственностью и моральными ориентирами, чтобы уважительно использовать подобные продукты, созданные с помощью технологий искусственного интеллекта.

В статье также обращается внимание на презентованную в 2018 году технологию Duplex американской транснациональной корпорации Google. По сути, это голосовой помощник, но с одной важной особенностью - его голос не отличается от голоса реального владельца программного обеспечения.

Задумка разработчиков состояла в том, чтобы освободить людей от рутинных операций, таких как заказы в магазинах и т. п., передать эти функции искусственному интеллекту. При этом возникло сразу несколько проблем. Такие голоса могут быть использованы для целенаправленного обмана людей с целью извлечения выгоды. Помимо этого, новая технология может привести к подрыву доверия при социальном взаимодействии.

- **Школа бизнеса имени Келли Университета Индианы[88].**
  В исследовании «Дипфейки: злой умысел или веселье» («Deepfakes: Trick or Treat?»)[89] содержатся следующие предложения по регулированию ДФ и борьбе с вредоносным ложным контентом: Фиксация исходного контента, гарантирующего выявление дипфейков. Эффективная борьба с дипфейками, созданными с целью обмана или манипулятивного воздействия на человека, предполагает систему раннего выявления и блокирования такого контента. В этой связи востребованными становятся технологии, в фоновом режиме отслеживающие и фиксирующие жизнь человека, включая его геолокацию, взаимодействие с другими людьми и учреждениями и прочее. С технической точки зрения сбор такой информации с современных мобильных устройств представляется достаточно простым и удобным. Однако, исходя из прямой угрозы неприкосновенности частной жизни и конфиденциальности действий человека, все собранные таким образом персональные данные должны быть сохранены, зашифрованы и использованы лишь в случае необходимости разоблачить дипфейк.

- **Разоблачение злонамеренных дипфейков.**
  Наряду с развитием технологий в сфере искусственного интеллекта, упрощающих процесс создания дипфейков, существуют также инновации, разработанные и для их обнаружения. Лидерами в этой области являются США. Например, Управление перспективных исследовательских проектов Министерства обороны США (Defense Advanced Research Projects Agency, DARPA) располагает программой судебной экспертизы цифрового контента СМИ, а также сервисами обнаружения дипфейков.

[88] www.kelley.iu.edu/
[89] www.researchgate.net/publication/338144721_Deepfakes_Trick_or_treat

Глобальные корпорации и цифровые платформы также вкладывает серьезные ресурсы в идентификацию ложного цифрового контента.

- **Защита прав в рамках закона.** Жертвы дипфейков должны иметь законодательно предусмотренную систему защиты от причиненного им материального или морального вреда в случае диффамации[90], злого умысла, нарушения неприкосновенности частной жизни или авторских прав, мошенничества или эмоционального расстройства.

- **Принятие мер по укреплению доверия к цифровой информации.** Глобальным цифровым компаниям, заявляющим о следовании общечеловеческим моральным принципам и правам человека, в контексте проблематики дипфейков следует предпринимать меры по укреплению доверия со стороны пользователей.

Среди инициатив, в повестку которых включены отдельные аспекты, связанные с дипфейками, можно отметить Партнерство по искусственному интеллекту (Partnership on AI)[91], независимую некоммерческую организацию «Общество будущего» (The Future Society)[92] и Институт инженеров электротехники и электроники (Institute of Electrical and Electronics Engineers)[93]. Также можно упомянуть и Глобальное партнерство по искусственному интеллекту (Global Partnership on Artificial Intelligence), в рамках которого страны уделяют внимание многосторонним формам сотрудничества по вопросам развития цифровых технологий, в том числе по ответственной разработке и законному использованию ИИ с учетом прав человека, инклюзивности, инноваций и экономического роста[94].

Считается, что проблема распространения вводящей в заблуждение цифровой информации может быть решена путем разработки технических стандартов сертификации источника медиаконтента. И сегодня этим уже занимается Коалиция за происхождение и подлинность контента (Coalition for Content Provenance and Authenticity, C2PA)[95]. C2PA — это проект Фонда совместного развития[96], сформированный в результате альянса компаний Adobe, Arm, Intel, Microsoft и Truepic.

[90] Диффамация — распространение порочащих сведений
[91] https://partnershiponai.org/
[92] www.thefuturesociety.org
[93] https://www.ieee.org// 1
[94] www.gpai.ai/
[95] www.c2pa.org/
[96] www.jointdevelopment.org/

4.2. Внутрикорпоративная политика и документы

Предполагается, что самый эффективный способ снизить негативные последствия дипфейков заключается не в ограничении их создания и публикации, а в регулировании онлайн-платформ, которые усиливают их негативный эффект[97].

При этом уже сегодня несколько популярных соцсетей и веб-сайтов официально запретили или планируют запретить экспансию вредоносных лживых материалов на своих площадках. К таковым относятся Facebook*, Instagram*, Twitter, Tiktok, PornHub и Reddit.

4.2.1. Twitter

В октябре 2020 года администрация сервиса микроблогов и социальной сети Twitter заявила, что вводит обновленную корпоративную политику[98] с целью адекватного противодействия дипфейкам и другому манипуляционному медиаконтенту, включающему в себя фото-, видео- или аудиофайлы, которые подверглись серьезным изменениям. С тех пор обнаруженный синтетический контент или манипуляционные файлы, вводящие в заблуждение, помечаются специальным маркером. Таким образом пользователи оказываются предупреждены раньше, чем успевают поделиться подобными материалами или лайкнуть их.

В рамках подготовки плана по работе с дипфейками специалисты Twitter предложили заинтересованным в решении проблемы акторам пройти опрос под хештегом #TwitterPolicyFeedback. На основе ответов соцсеть планировала составить репрезентативное мнение о том, следует ли удалять измененные фотографии и видео, обойтись лишь предупреждающими надписями или вовсе никак не ограничивать такой контент. В итоге было принято решение, что материалы будут удаляться автоматически, если они угрожают чьей-либо физической безопасности или чьему-либо психическому здоровью, ущемляют чье-либо достоинство или нарушают чью-либо неприкосновенность частной жизни.

4.2.2. Meta* (Facebook*, Instagram*)

Американская транснациональная холдинговая компания Meta* стремится удалять дипфейки или иным образом измененные медиафайлы, где манипуляция неочевидна и может ввести в заблуждение, особенно в случаях с видеоконтентом[99].

Обновленная корпоративная политика Facebook*, принятая в январе 2020 года, ввела запрет на распространение видео, которые были отредактированы или синтезированы таким образом, что обычным пользователям непросто распознать подделку[100]. При этом соцсеть отказалась запрещать развлекательный пародийный или сатирический видеоконтент. В этой связи также стоит отметить, что новые правила никак не ограничивают более простые формы введения пользователей в заблуждение, такие как неправильная маркировка материалов, склеивание диалогов или выдергивание цитат из контекста.

[97] https://law.unh.edu/sites/default/files/media/2022/06/kryskowski_pp114-185.pdf
[98] *21 марта 2022 года Тверской районный суд признал организацию Meta (социальные сети Instagram и Facebook) экстремистской, тем самым запретив ее деятельность в России www.techcrunch.com/2019/11/11/twitter-drafts-a-deepfake-policy-that-would-label-and-warn-but-not-remove-manipulated-media/
[99] https://transparency.fb.com/en-gb/policies/community-standards/manipulated-media/
[100] www.about.fb.com/news/2020/01/enforcing-against-manipulated-media/

Кроме того, в структуре Meta* существует Наблюдательный совет из представителей различных культур и профессий. Он рассматривает обращения пользователей в связи с их недовольством тем или иным заключением компании в отношении определенного контента в Facebook* и Instagram*. Совет стремится продвигать свободу слова и принимать независимые решения[101].

Также в 2019 году социальная сеть Facebook* объединилась с компанией Microsoft и учеными из некоммерческой коалиции Partnership on AI, которая привержена ответственному использованию искусственного интеллекта. Посредством конкурса Deepfake Detection Challenge стороны добились создания решений по обнаружению дипфейков[102], а в последствии Facebook* стал активно к ним прибегать. Однако, стоит отметить, что пока эти алгоритмы выявляют только две трети ложной информации[103].

4.2.3. Snapchat/TikTok

Несколько иной подход к феномену дипфейков демонстрируют платформы Snapchat и TikTok, которые усматривают в них не только риски, но и новые возможности по работе с аудиторией. Так, в начале 2020 года владельцы мобильного приложения обмена сообщениями Snapchat приобрели компанию AI Factory, с которой ранее разрабатывали функцию Cameo, позволяющую людям вставлять свое лицо в уже готовые шаблоны. Такой шаг, как полагают наблюдатели, означал, что Snapchat решил и дальше совершенствовать свои возможности интерактивного погружения пользователей в различный видеоконтент[104].

На этом фоне сервис для создания и просмотра коротких видео TikTok заявил о намерении продолжать развитие своей встроенной функции по оцифровке внешности пользователей. Она подразумевает биометрическое сканирование лица человека с разных углов с последующим его наслоением на любое видео из каталога платформы, в том числе в различные сцены фильмов.

В связи с рассматриваемой проблемой стоит отметить, что в 2020 году TikTok также утвердил политику, запрещающую синтетический или манипуляционный контент, который может нанести вред пользователям или широкой общественности. В частности, Принципы сообщества TikTok запрещают обмен любого рода данными, вводящими в заблуждение относительно выборов или других гражданских процессов, распространяемыми в рамках кампаний по дезинформации, и прочей ложной информацией[105].

[101] www.oversightboard.com/
[102] www.reuters.com/article/us-facebook-microsoft-deepfakes/facebook-microsoft-launch-contest-to-detect-deepfake-videos-idUSKCN1VQ2T5
[103] www.wired.com/story/deepfakes-not-very-good-nor-tools-detect/
[104] www.socialmediatoday.com/news/snapchat-and-tiktok-are-both-reportedly-working-on-new-deepfake-type-feat/569792/

4.2.4. YouTube

На видеохостинге YouTube действует запрет на манипуляционные медиа в соответствии с Правилами в отношении ложной информации[106]. На платформе недопустимо распространение контента, способного нанести серьезный вред, в том числе в реальной жизни. Например, запрещены видео, в которых продвигаются вредные лекарства и методы лечения, а также ролики, сфальсифицированные с помощью технических средств, и материалы, препятствующие демократическим процессам.

4.2.5. Microsoft

По оценке одной из крупнейших транснациональных компаний по производству проприетарного программного обеспечения для различного рода вычислительной техники Microsoft в долгосрочной перспективе ни люди, ни методы искусственного интеллекта не смогут надежно отличить факты от синтезируемых вымыслов. Поэтому миру необходимо срочно подготовиться к ожидаемой опасной траектории — появлению все более реалистичных и убедительных дипфейков[107]. Противодействовать угрозе, по мнению специалистов корпорации, способны технологии идентификации происхождения цифрового контента.

В борьбе с дипфейками, которые могут использоваться для распространения дезинформации, исследовательское подразделение компании Microsoft Research, Управление ответственного искусственного интеллекта (The Office of Responsible AI) и Комитет по эфиру (The Aether Committee) централизованно разработали инструмент Microsoft Video Authenticator, который использует специальный алгоритм для определения подлинности фотографий или видеороликов[108].

[105] www.tiktok.com/community-guidelines?lang=en
[106] www.support.google.com/youtube/answer/10834785
[107] https://blogs.microsoft.com/on-the-issues/2022/05/03/artificial-intelligence-department-of-defense-cyber-missions/
[108] https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/

4.3. Рассматриваемые и принятые нормативно-правовые акты, другие инструменты регулирования, правоприменительные практики

Правительства многих стран ведут поэтапную работу по совершенствованию своих правовых систем на предмет дипфейков. Наиболее активно, в том числе и на законодательном уровне, противодействуют распространению вредоносного ложного контента Индия, КНР, США, Сингапур, Великобритания, Южная Корея, Япония, Австрия и Евросоюз. Австралия планирует запретить дипфейки в большинстве штатов, где действуют законы о сексуальном насилии на основе изображений[109].

Однако стоит отметить, что разработка подобных документов требует серьезной экспертной дискуссии и всестороннего анализа, поэтому, как правило, в целом во всем мире затягивается по времени. Процесс принятия в странах необходимых законодательных мер реализуется пока достаточно медленно еще также и потому, что проблема дипфейков многослойная, пронизывает ряд других сфер, затрагивает острые и чувствительные вопросы. Кроме того, требуемые нововведения весьма непросто решить на техническом уровне. В связи со всем этим пока что все-таки больше наблюдается курс на внедрение национальных законодательных мер, именно нацеленных на возложение большей ответственности за создание и распространение ложного контента на онлайн-платформы.

4.3.1. Европейский союз

Европейский союз продолжает предпринимать шаги по пресечению возможных злоупотреблений в сфере искусственного интеллекта. Так, с 2018 года реализуется Стратегия Европейской комиссии по ИИ (European Commission Artificial Intellegence Strategy)[110], которая, в частности, прописывает принцип контроля человека над искусственным интеллектом в вопросах регулирования разработок с его использованием.

Кроме того, необходимость повышения осведомленности пользователей о новейших технологиях, в том числе о дипфейках, упоминается в Плане действий Евросоюза по дезинформации от декабря 2018 года (Action Plan against Disinformation)[111].

Также в феврале 2020 года Европейская комиссия опубликовала «Белую книгу искусственного интеллекта» (White Paper on AI)[112], в которой одновременно излагаются варианты по достижению внедрения ИИ и устранению рисков, связанных с его использованием.

В апреле 2021 года для развития экосистемы доверия в проект Регламента ЕС о гармоничных правилах в отношении искусственного интеллекта были внесены новые предложения[113]. К примеру, были рекомендованы минимальные обязательства по прозрачности для систем, взаимодействующих с людьми.

В Резолюции от января 2021 года[114] Европейский парламент призывал к обязательной маркировке дипфейков.

[109] www.wkls.com.au/deepfake-technology-current-remedies-and-possible-legal-consequences/#_ftn9
[110] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&from=EN
[111] https://www.eeas.europa.eu/sites/default/files/action_plan_against_disinformation.pdf
[112] https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en
[113] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206
[114] https://www.europarl.europa.eu/doceo/document/TA-9-2021-0009_EN.html

В документе сказано, что все синтетические материалы должны быть помечены создателем как неоригинальные, применены в рамках правовых норм и не должны быть использованы в избирательных кампаниях.

Также в январе 2021 года в Отчете об искусственном интеллекте[115] Европарламент призвал к проведению исследований в данной сфере. По их завершении практики и технологии противодействия не должны отставать от приемов злонамеренного использованием ИИ.

В тексте Закона Евросоюза о цифровых услугах от 19 октября 2022 года (The Digital Services Act, DSA)[116] изложены правила фиксации и удаления незаконного контента. Они, среди прочего, накладывают на крупные онлайн-платформы обязательство по проведению оценки рисков. В частности, речь идет о преднамеренном манипулировании цифровыми услугами с негативным влиянием на защиту общественного здоровья, несовершеннолетних, гражданское общество, выборы, безопасность и т. д. При этом за невыполнение требований DSA, в т. ч. мер противодействия дипфейкам и поддельным учетным записям на своих платформах, компании могут быть оштрафованы на сумму до шести процентов от их годового глобального оборота в предыдущем финансовом году.

В 2022 году в Евросоюзе был принят Усиленный кодекс по борьбе с дезинформацией (The Strengthened Code of Practice on Disinformation)[117], в котором в частности определены обязательства подписавших его организаций по снижению воздействия ложного цифрового контента. В соответствии с документом подписанты должны иметь прозрачную политику в отношении алгоритмов для обнаружения и модерации дипфейков.

По состоянию на 16 июня 2022 года кодексу уже следовали 34 крупные компании, среди которых — Meta*, Twitter, Google, Microsoft[118]. А в июле того же года Еврокомиссия опубликовала официальный Призыв к подписанию Усиленного кодекса по борьбе с дезинформацией[119]. Согласно ему, отраслевые компании должны придерживаться стандартов саморегулирования для борьбы с дезинформацией.

Кроме того, стоит отметить, что программа Horizon 2020[120] запустила грантовый проект WeVerify[121] для сложных задач проверки контента. В его рамках цифровая среда анализируется на предмет наличия дезинформации, распространяемой в том числе посредством дипфейков. Для этого применяется как целевое обнаружение, так и технология блокчейн. Сегодня финансирование проекта продолжает программа Horizon Europe[122].

По мнению экспертов Европейского парламента, любые ограничения должны быть сбалансированы со свободой выражения мнения, творчества и созидания в науке[123]. Однако возможность создания дипфейков не должна автоматически давать права на их широкое распространение.

[115] www.europarl.europa.eu/doceo/document/A-9-2021-0001_EN.html
[116] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en
[117] www.regmedia.co.uk/2022/06/16/eu_code_of_practice_2022.pdf
*21 марта 2022 года Тверской районный суд признал организацию Meta (социальные сети Instagram и Facebook) экстремистской, тем самым запретив ее деятельность в России
[118] https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation
[119] https://digital-strategy.ec.europa.eu/en/news/call-interest-become-signatory-2022-code-practice-disinformation
[120] https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-2020_en
[121] https://weverify.eu/#
[122] https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe_en
[123] https://www.europarl.europa.eu/RegData/etudes/ATAG/2021/690046/EPRS_ATA(2021)690046_EN.pdf

Важно учитывать роль вредоносных ДФ в социальных и политических процессах. В этом контексте в Резолюции от 19 мая 2021 года[124] об искусственном интеллекте в образовании, культуре и аудиовизуальном секторе Европарламент подчеркнул большое значение плюрализма в СМИ, качественной журналистики и общего уровня осведомленности.

В отчете Европола «Это реальность? Правоприменение и проблема дипфейков» («Facing reality? Law enforcement and the challenge of deepfakes») 2022 года[125] ДФ были определены как опасность для граждан Евросоюза, а также был приведен список актуальных рисков. Согласно документу, противодействие синтетическому контенту потребует от правительств скоординированной адаптации всех действенных инструментов.

4.3.2. Республика Индия

В Индии пока не создана правовая база регулирования оборота или запрета дипфейков. Действующее законодательство позволяет в необходимых случаях ограничивать их использование и распространение.

Процедуры удаления нежелательного контента реализуются в соответствии с текущими нормами национального законодательства, в частности с положениями разделов № 67 и № 67A Закона об информационных технологиях 2000 года[126]. Ими предусмотрена ответственность за публикацию в электронном виде материалов откровенно сексуального характера. Кроме того, раздел № 500 Уголовного кодекса Индии[127] предполагает наказание за диффамацию.

Право на неприкосновенность частной жизни является одним из основных в национальном законодательстве Индии. В этой связи определенные надежды в стране связывались с Законопроектом о защите персональных данных 2019 года[128]. Он в том числе предусматривал защиту информации, по которой можно прямо или косвенно идентифицировать человека. Документ накладывал ограничения на обработку таких данных за исключением четко определенных законом случаев[129]. Но в августе 2022 года законопроект был отозван с пояснением, что будет представлена его более полная версия.

Одобренные в 2021 году Руководящие принципы для посредников и Кодекс этики цифровых СМИ[130] (Intermediary Guidelines and Digital Media Ethics Code) обязывают посредников в течении 24 часов принимать меры по удалению контента или отключению доступа к нему в тех случаях, если в нем кто-то выдает себя за другого человека и если он создан с использованием искусственных нейронных сетей и методов глубокого обучения[131].

4.3.3. Китайская Народная Республика

1 марта 2022 года в Китае вступили в силу Положения об управлении алгоритмическими рекомендациями для информационных интернет-услуг (Provisions on the Management of Algorithmic Recommendations for Internet Information Services), или так называемые Положения об алгоритмах. В частности, в них содержатся общие требования к маркировке синтетической информации и отдельные меры пресечения вредоносного контента.

[124] https://www.europarl.europa.eu/doceo/document/TA-9-2021-0238_EN.html
[125] https://www.europol. europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf
[126] https://eprocure.gov.in/cppp/rulesandprocs/kbadqkdlcswfjdelrquehwuxcfmijmuixngudufgbuubgubfugbububjxcgfvsbdihbgfGhdfgFHytyhRtMTk4NzY=
[127] https://legislative.gov.in/sites/default/files/A1860-45.pdf
[128] https://164.100.47.193/lob/17/IX/SLOB3.8.2022.pdf
[129] https://blogs.lse.ac.uk/southasia/2020/05/21/deepfakes-in-india-regulation-and-privacy/
[130] https://mib.gov.in/digital-media-guidelines-and-policies
[131] https://164.100.47.194/Loksabha/Questions/QResult15.aspx?qref=27009&lsno=17

Ниже выборочно приведены выдержки из документа[132]:

- «поставщики услуг алгоритмических рекомендаций должны… пополнять базы данных признаков, которые будут использоваться для выявления незаконной и вредной информации, совершенствовать стандарты и процессы пополнениях таких баз»;
- «при обнаружении противоправной информации следует немедленно прекратить ее распространение, принять меры по удалению, а также сообщить об этом в соответствующие ведомства»;
- «поставщики услуг алгоритмических рекомендаций… не могут генерировать или синтезировать фальшивые новости, а также распространять новостную информацию, не опубликованную профильными подразделениями в установленном государством порядке».

28 января 2022 года в Китае был опубликован проект Положений об управлении глубоким синтезом в информационных интернет-сервисах (Provisions on the Administration of Deep Synthesis Internet Information Services)[133]. Создатели документа отнесли к глубокому синтезу (deep synthesis technology) технологии, использующие генеративные алгоритмы (generative sequencing algorithms) для создания контента. Среди таковых:

- технологии для создания или редактирования текстового контента, такие как генерирование глав, преобразование стиля текста;
- технологии для создания или редактирования голосового контента, такие как преобразование текста в речь и изменение голоса;
- технологии для создания или редактирования неголосового аудиоконтента, например для написания музыки;
- технологии для создания или редактирования биометрических признаков — вставка или редактирование лиц или жестов в фото или видео, генерирование внешности;
- технологии для редактирования небиометрических признаков — улучшение или восстановление изображений и видео;
- технологии для создания или редактирования виртуальных сцен — 3D-реконструкция.

Документ затрагивает и вопросы маркировки синтетического контента для оповещения пользователей. Среди подлежащих маркировке содержатся следующие технологии глубокого синтеза:

- виртуальное общение с ботом в виде реального человека для создания или редактирования текстов;
- генерирование речи, имитация или редактирование голоса;
- создание или редактирование лица или жестов;
- генерирование иммерсивных сцен.

Положения об управлении глубоким синтезом в информационных интернет-сервисах могут рассматриваться как логическое продолжение двух других инструментов, принятых и рассматриваемых в Китае ранее.

[132] При цитировании Положений об алгоритмах использовался английский текст с сайтов DigiChina и China Low Translate: https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022/; https://www.chinalawtranslate.com/en/algorithms/. Перевод на английский язык с китайского на указанных ресурсах не является официальным
[133] При описании/цитировании документа использован английский текст с сайта China Low Translate: https://www.chinalawtranslate.com/en/deep-synthesis-draft/. Перевод на английский язык с китайского на указанном ресурсе не является официальным

Речь идет о Правилах администрирования онлайновых аудио- и видеоинформационных услуг от 1 января 2020 года (Regulations on the Administration of Online Audio and Video Information Services[134] и проекте Положения об управлении производством радио-, телевизионного и онлайн-контента от 2019 года (Provisions on Management of Online A/V Information Services, National Radio and Television Administration)[135].

Первый документ запрещает использование синтетических изображений, аудио- и видеоматериалов для распространения недостоверной информации и содержит 19 статей со следующими основными положениями:

- поставщик онлайновых аудио- и видеоинформационных услуг должен проверять подлинность идентификационных данных пользователей;
- ни одна организация и ни одно частное лицо не должны использовать онлайновые аудио- и видеоинформационные услуги и связанные с ними информационные технологии для участия в незаконной деятельности и/или публикации незаконного контента;
- если поставщик услуг и его пользователи применяют новые технологии на основе глубокого обучения и виртуальной реальности для производства и распространения нереальной аудио- и видеоинформации, они должны маркировать ее соответствующим образом;
- поставщик онлайновых аудио- и видеоинформационных услуг должен создать механизм опровержения слухов.

Стоит отметить, что рядом экспертов высказывается мнение о том, что с учетом разработки обозначенных выше документов Китай опережает Евросоюз и США в части регулирования синтетического контента[136].

С 10 января 2023 года в стране действуют новые правила для интернет-провайдеров, призванные обеспечить защиту граждан от использования их изображений без согласия в дипфейках[137].

4.3.4. Республика Сингапур

В Сингапуре отсутствуют законы, касающиеся непосредственно дипфейков. При этом отдельные нормативно-правовые акты могут частично снять некоторые проблемы в этой области.

Так, Закон о защите ото лжи и манипуляций в Интернете от 2019 года (Protection from Online Falsehoods and Manipulation Act)[138] позволяет удалять дезинформацию, которая угрожает национальным интересам. В этой связи он также может быть использован для устранения дипфейков[139].

Кроме того, в разделе № 499 Уголовного кодекса Сингапура[140] говорится: если гражданин сознательно намеревается опорочить другого человека, то он совершает преступление. Это положение, вероятно, может действовать и при использовании дипфейков с целью оклеветать других лиц.

Эксперты считают, что законы о сексуальных домогательствах тоже могут быть применены для защиты людей от вредных последствий

[134] При описании/цитировании документа использован английский текст с сайта China Low Translate: https://www.chinajusticeobserver.com/law/x/online-audio-and-video-information-services-20191118
Перевод на английский язык с китайского на указанном ресурсе не является официальным
[135] При описании/цитировании документа использован английский текст с сайта China Low Translate: https://www.chinalawtranslate.com/en/draft-program-management/.
Перевод на английский язык с китайского на указанном ресурсе не является официальным
[136] https://www.insideprivacy.com/artificial-intelligence/china-takes-the-lead-on-regulating-novel-technologies-new-regulations-on-algorithmic-recommendations-and-deep-synthesis-technologies/
[137] www.reuters.com/technology/chinas-rules-deepfakes-take-effect-jan-10-2022-12-12/
[138] https://www.pofmaoffice.gov.sg/regulations/protection-from-online-falsehoods-and-manipulation-act/
[139] https://spj.hkspublications.org/2021/12/11/deepfakes-the-implications-of-this-emerging-technology-on-society-and-governance/#_edn12
[140] https://www.ilo.org/dyn/natlex/docs/ELECTRONIC/67824/65070/F-560310733/SGP67824%202015.pdf

дипфейков, а законы об авторском праве способны предотвратить преступное использование интеллектуальной собственности в них. Например, раздел № 377BE Уголовного кодекса Сингапура наказывает за распространение или угрозу распространения интимного изображения или записи лица[141]. При этом с авторским правом не все так однозначно. Так, Ведомство интеллектуальной собственности Сингапура (Intellectual Property of Singapore) признает, что законов об интеллектуальной собственности недостаточно для борьбы с дипфейками — должен быть междисциплинарный подход, включающий уголовное и гражданское права[142].

4.3.5. Соединенное Королевство Великобритании и Северной Ирландии

Сегодня в Великобритании нет законов, направленных на борьбу с угрозами дипфейков. Также нет и каких-либо отдельных положений, касающихся прав интеллектуальной собственности применительно к синтетическому контенту, на которые можно было бы ссылаться в судебных спорах. Но стоит отметить, что британское правительство признает важность изучения этого вопроса, в частности в контексте уголовного права[143].

Так, в 2018 году власти поручили Правовой комиссии Великобритании изучить дипфейки в рамках проблем, связанных с порнографией. После анализа ситуации она рекомендовала пересмотреть меры уголовного законодательства за нарушение конфиденциальности в Интернете, а также учесть наказания за вред, наносимый сексуализированными дипфейками[144].

4.3.6. Соединенные Штаты Америки

США предпринимают значительные меры для предотвращения информационных угроз и обеспечения ситуационного контроля в условиях внутренних и внешних вызовов дипфейков. Комиссия национальной безопасности по искусственному интеллекту относит ДФ к новым видам опасностей, исходящим от ИИ-систем[145].

Если вдаваться в конкретику, то Закон США «Об определении результатов генеративных состязательных сетей» 2020 года (Identifying Outputs of Generative Adversarial Networks Act)[146] требует от Национального научного фонда исследовать дипфейки, а от Национального института стандартов и технологий — оказывать поддержку в создании типовых образцов работы с синтетическим контентом. Кроме того, согласно тексту документа, оба органа должны предлагать форматы взаимодействия с бизнесом по вопросу идентификации ДФ.

Закон о национальной обороне на 2020 финансовый год (National Defense Authorization Act for Fiscal Year 2020)[147] учредил премию Deepfakes, цель которой — содействие актуальным исследованиям и разработкам в этой сфере, а также коммерциализация продуктов для обнаружения синтетического контента.

Закон о национальной обороне на 2021 финансовый год (National Defense Authorization Act for Fiscal Year 2021)[148] ввел требование к Министерству внутренней безопасности США рассматривать ситуацию с дипфейками в ежегодных отчетах в течение следующих пяти лет[149].

[141] www.lawtech.asia/fake-porn-real-harm-examining-the-laws-against-deepfake-pornography-in-singapore/
[142] https://www.wipo.int/export/sites/www/about-ip/en/artificial_intelligence/call_for_comments/pdf/ms_singapore.pdf
[143] www.natlawreview.com/article/rise-deepfake-demands-urgent-legal-reform-uk
[144] www.natlawreview.com/article/rise-deepfake-demands-urgent-legal-reform-uk
[145] www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf
[146] www.govtrack.us/congress/bills/116/s2904
[147] www.congress.gov/bill/116th-congress/senate-bill/1790/text?q=%7B%22search%22%3A%5B%22ndaa%22%5D%7D&r=2&s=1
[148] www.congress.gov/116/plaws/publ283/PLAW-116publ283.pdf
[149] www.asisonline.org/security-management-magazine/latest-news/today-in-security/2021/january/U-S-Laws-Address-Deepfakes/

В докладах должны охватываться все виды потенциального вреда технологии — от кампаний иностранного влияния до мошенничества.

Закон о государственной обороне на 2023 финансовый год (National Defense Authorization Act for Fiscal Year 2023)[150] также включает статьи, направленные на решение растущей проблемы дипфейков. Например, документ требует проводить разведывательную оценку угроз, которые исходят от иностранных правительств и негосударственных акторов, создающих или использующих машинно-манипулируемые медиа с участием военнослужащих и членов их семей.

Также Управление перспективных исследовательских проектов Министерства обороны США с целью предотвращения размещения ложной информации и распознавания дипфейков запустило государственную программу анализа медиаматериалов. В ее рамках реализуются проекты MediFor (Media Forensics)[151] и SemaFor (Semantic Forensics).

Цель MediFor — разработка алгоритмов автоматической оценки целостности фотографий и видео, а также сбора информации о способах создания лживого контента. А цель SemaFor — создание алгоритмов машинного обнаружения различных типов дипфейков.

Кроме того, в 2021 году было заявлено, что американские военные исследователи разработали новый метод обнаружения дипфейков. Он позволит создать современную технологию Soldier для решения важных задач, таких как распознавание угроз противника[152].

Несколько штатов США, например Виргиния, приняли локальные законы о наказаниях за порнографические дипфейки. Штат Техас при этом дополнительно причислил к правонарушениям создание синтетического контента с целью влияния на электоральный процесс. Вместе с тем штаты Массачусетс и Калифорния так и не смогли одобрить законопроекты, направленные на противодействие дипфейкам, чтобы не создавать избыточную нормативную базу в ИТ-отрасли[153].

Еще один рассматриваемый сегодня в США законопроект — «Об ответственности за дипфейки» (Deepfakes Accountability Act) от 2019 года[154]. Он устанавливает требования к поддельному контенту, вводит уголовные наказания за связанные с ним нарушения, а также предписывает формирование в Министерстве внутренней безопасности США целевой группы, которая станет ядром участия правительства в практике создания дипфейков и любых контрмер для борьбы с ними.

Также в 2021 году был внесен на рассмотрение смежный законопроект для решения проблемы дезинформации в США — Deepfake Task Force Act[155]. Его основная цель — создание все той же целевой группы по дипфейкам. Она, как предполагается, должна состоять из представителей правительства, бизнеса и научных кругов.

В Федеральном бюро расследований США (FBI) действует Центр жалоб на интернет-преступления (Internet Crime Complaint Center)[156]. В него могут обратиться все, кто

[150] https://www.congress.gov/bill/117 th-congress/house-bill/7900/text?format=txt
[151] www.grfc.ru/grfc/news/detail/index.php?ID=49515
[152] www.independent.co.uk/life-style/gadgets-and-tech/us-army-deepfake-detection-tool-b1840217.html
[153] https://www.jdsupra.com/legalnews/first-federal-legislation-on-deepfakes-42346/
[154] https://www.congress.gov/bill/116 th-congress/house-bill /3230
[155] https://www.congress.gov/bill /117 th-congress/senate-bill /2559/ text
[156] https://www.ic 3.gov/

считает себя жертвой интернет-преступлений[157]. Также есть возможность подать жалобу на дипфейк или онлайн-кражу персональной информации[158].

4.3.7. Южная Корея

В июне 2020 года Южная Корея пересмотрела закон «Об особых делах, касающихся наказания за сексуальные преступления» (Act on Special Cases Concerning the Punishment of Sexual Crimes)[159]. В частности, согласно внесенным поправкам, создатели дипфейковых видео, которые сделаны без одобрения их участников и способны вызвать сексуальное желание или оскорбить, должны быть приговорены к тюремному заключению на срок до пяти лет или к штрафу в размере до 50 млн вон. При этом если преступление было совершено с целью получения коммерческой выгоды, то тюремный срок должен быть увеличен до семи лет[160].

Кроме того, в Южной Корее функционирует База данных для обнаружения дипфейков (Korean Deepfake Detection Dataset)[161]. В ней собраны синтетические и реальные видео. Ее цель — помочь специалистам в разработке инновационных методов обнаружения ДФ.

4.3.8. Япония

В соответствии со статьей № 175 Уголовного кодекса Японии[162], гражданин, который демонстрирует на публику непристойности, наказывается лишением свободы на срок до двух лет и/или штрафом в размере до 2,5 млн иен. То же самое относится к лицам, которые распространяют непристойные записи, в том числе электронные и магнитные, посредством телекоммуникации. В ноябре 2021 года в Японии впервые по уголовному делу был арестован создатель порнографического дипфейк-контента[163].

Министерство иностранных дел Японии планирует запустить в 2023 году систему для сбора и анализа фейковой информации в соцсетях и на других контент-платформах, основанную на технологиях ИИ[164].

4.3.9. Австрия

Австрийское правительство придерживается позиции о том, что регулирование дипфейков должно учитывать основные права человека, особое внимание следует уделить защите свободы выражения мнения и свободы творчества[165].

В мае 2022 года правительство страны опубликовало План действий по борьбе с дипфейками (Aktionsplan Deepfake)[166]. Для достижения его целей было организовано активное межведомственное и общеевропейское сотрудничество. Документ, в частности призвал онлайн-платформы стать более подотчетными в отношении ДФ.

Сегодня на основании плана ведется исследовательская работа и создается необходимое программное обеспечение[167].

[157] www.ic3.gov/Home/FileComplaint
[158] www.ic3.gov/Media/Y2022/PSA220628
[159] https://elaw.klri.re.kr/eng_service/lawView.do?hseq=40947&lang=ENG
[160] https://en.yna.co.kr/view/AEN20210114006500315
[161] https://deepbrainai-research.github.io/kodf/
[162] www.japaneselawtranslation.go.jp/en/laws/view/3581/en#je_pt2ch24at2
[163] www.vice.com/en/article/xgdq87/deepfakes-japan-arrest-japanese-porn
[164] https://asia.nikkei.com/Business/Technology/Japan-tap
[165] www.euractiv.com/section/disinformation/news/austria-to-combat-deep-fakes-amid-increasing-use-of-the-technology/
[166] www.bmi.gv.at/bmi_documents/2779.pdf
[167] www.parlament.gv.at/PAKT/PR/JAHR_2022/PK1005/#

5. Возможности совершенствования инструментов регулирования в России

В России набор инструментов регулирования дипфейков находится на этапе формирования.

В ноябре 2018 года был заключен Меморандум о сотрудничестве в сфере охраны исключительных прав в эпоху развития цифровых технологий[168]. В числе прочего он предусматривает обязательства операторов поисковых систем прекратить выдачу ссылок на нелегально размещенные аудиовизуальные произведения правообладателей по их заявлениям.

Под документом поставили подписи представители 12 компаний. Среди них — Первый канал, Всероссийская государственная телевизионная и радиовещательная компания, «СТС Медиа», «Газпром-Медиа Холдинг», Национальная медиагруппа, Ассоциация продюсеров кино и телевидения, Ассоциация по стимулированию оборота легального контента в Интернете, «Яндекс», Mail.Ru, Rambler & Co, «Руформ» и «Кинопоиск». Меморандум продлевается на ежегодной основе[169].

[168] www.roem.ru/wp-content/uploads/2018/11/2018.11.01.itog.memorandum.pravki.po.pdf
[169] www.np-mks.com/press-tsentr/novosti/dejstvie-antipiratskogo-memoranduma-prodleno-na-god.html

Сегодня в России действует порядок ограничения доступа к информации, распространяемой с нарушением правил. Это определяется в статье № 15.3 Федерального закона об информации, информационных технологиях и защите информации от 27 июля 2006 года[170].

Кроме того, публичное распространение заведомо ложной информации об обстоятельствах, представляющих угрозу жизни и безопасности граждан, и публичное распространение заведомо ложной общественно значимой информации, повлекшее тяжкие последствия, наказываются согласно статье № 207.1 Уголовного кодекса Российской Федерации от 13 июня 1996 года[171]. За распространение фейковых новостей предусмотрена ответственность согласно статье № 13.15 Кодекса Российской Федерации об административных правонарушениях[172].

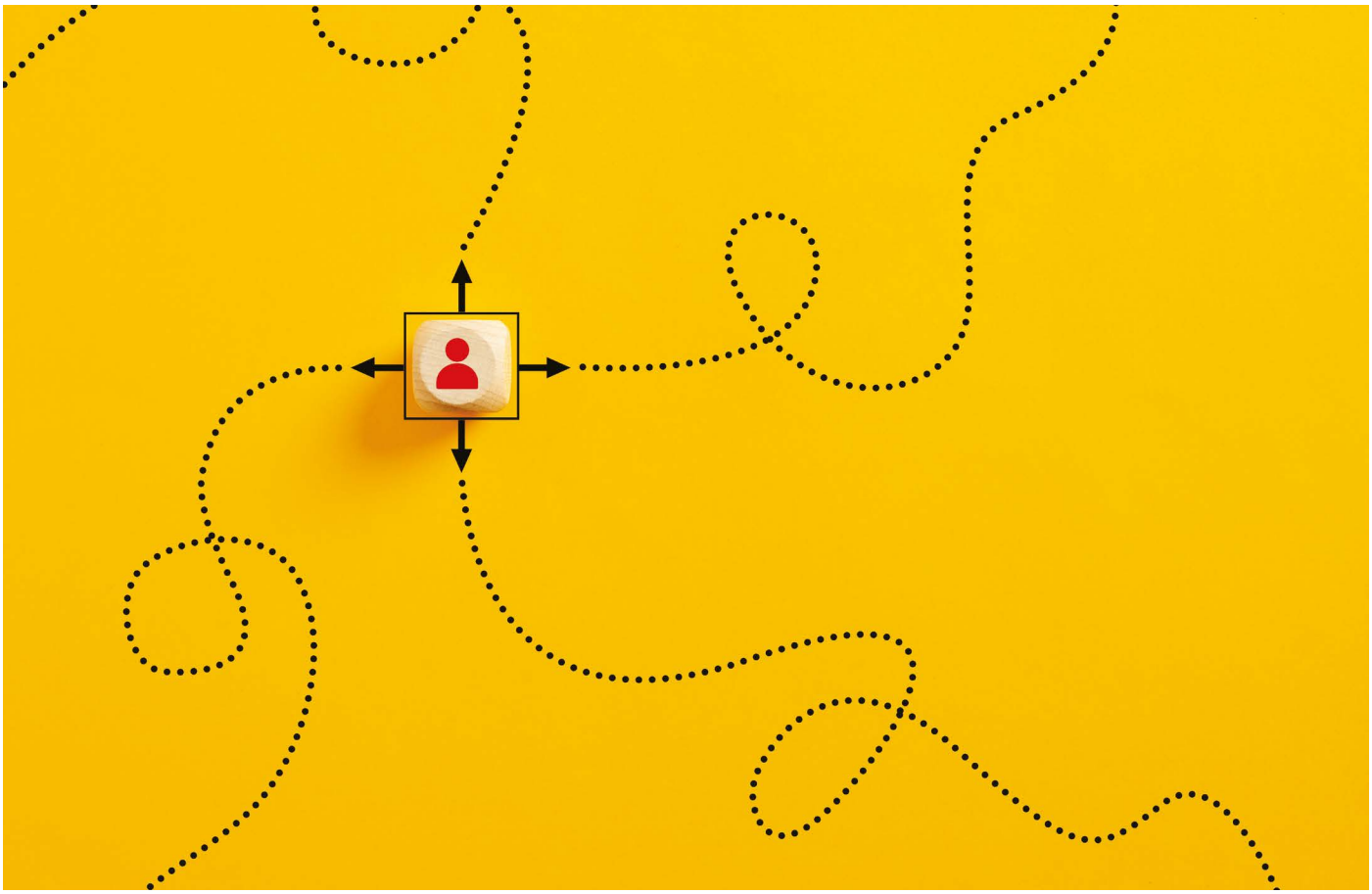
В стране также предпринимаются усилия по совершенствованию технических возможностей выявления дипфейков различной природы. Так, в 2021–2022 годах по заказу Министерства внутренних дел Российской Федерации проведена научная работа «Исследование возможных способов выявления признаков внутрикадрового монтажа видеоизображений, выполненного с помощью нейронных сетей»[173]. Контракт на ИТ-разработку был заключен под шифром «Зеркало» (также используется название «Верблюд»). Проект предполагает разработку техзадания по созданию аппаратно-программного комплекса для выявления дипфейков, содержащихся в видеофайлах распространенных форматов, в том числе из сети Интернет. В МВД России считают, что инновация повысит уровень научно-технического обеспечения работы экспертно-криминалистических подразделений при проведении видеотехнических экспертиз.

В марте 2022 года ученые Федерального исследовательского центра «Информатика и управление» Российской академии наук[174] и Московского физико-технического института[175] разработали метод детектирования подделок (спуфинга) в мобильных системах распознавания по лицу с помощью штатной стереокамеры[176]. По сравнению с аналогами он работает быстрее, что позволит применять его в биометрических системах.

В августе 2022 года ПАО Сбербанк разработал и запатентовал технологии по распознаванию дипфейков. Объединенные в одну систему, они позволяют с высокой точностью определить синтетически измененные изображения лиц на видео[177].

[170] www.consultant.ru/document/cons_doc_LAW_61798/34547c9b6ddb60cebd0a67593943fd9ef64ebdd0/
[171] www.consultant.ru/document/cons_doc_LAW_10699/9d8a5b6501a01da934c1bbd0ca9b1fd46df76a72/
[172] www.consultant.ru/document/cons_doc_LAW_34661/82c0a663173b440cc9b027bc8e687dc9e36e71ad/
[173] https://www.tadviser.ru/index.php/Проект:Зеркало_(Верблюд)_-_система_МВД_для_распознавания_фейковых_видео
[174] www.frccsc.ru
[175] www.mipt.ru/
[176] www.ras.ru/news/shownews.aspx?id=104d40b1-67ab-4c30-b7c1-42958ac4b457&print=1
[177] www.ria.ru/20220817/dipfeyki-1810194087.html?ysclid=lc5zcmvwap357509370

6. Общие выводы и рекомендации



Спектр незаконного использования дипфейков, а также их применение с негативными для общества последствиями достаточно широк. С учетом быстрого развития технологий будут разрабатываться все новые формы ДФ, что увеличит масштабы различных видов ущерба.

На данном этапе важно объективно исследовать риски и предотвратить демонизацию ДФ-технологий, не нанеся ущерба их применению в полезных целях.

Борьба с вредоносными дипфейками требует высокого уровня технических знаний и во многом зависит от разработки специальных новейших методик, программного обеспечения и наличия соответствующего оборудования.

Меры противодействия вредоносным дипфейкам можно классифицировать по следующим широким категориям:

- законодательные акты и правоприменение;
- политики, подходы и решения по управлению экосистемами, платформами, крупными ресурсами;
- инженерно-технические меры, технологические решения и проактивное позитивно направленное вмешательство на аппаратном и программном уровнях: мониторинг, выявление, блокировка, удаление;
- документы мягкого права и саморегулирование, повышение ответственности акторов;
- наращивание информированности и медиаграмотности населения, пропаганда мер цифровой гигиены, различные меры системной профилактики в среде акторов и пользователей.

Вышеперечисленные задачи ориентированы на государственные профильные ведомства и службы, правовые институты, научно-исследовательские центры и организации, в том числе по разработке ПО, владельцев платформ и экосистем, частные технологические и сервисные компании.

Немаловажным будет фактор наличия у профильных ведомств и организаций инструментов выявления современных зарубежных дипфейков, производимых на основе последних технологических достижений. Это, в частности, предполагает системный сбор и анализ информации о состоянии иностранных разработок в данной сфере. При этом особое внимание целесообразно обратить на такие виды дипфейков, которые способны нанести явный или скрытый ущерб национальной безопасности России, а также ДФ-продукты с кумулятивным эффектом, который наступает при накоплении их критической массы в определенном медиапространстве или цифровом медиасегменте страны или региона.

В ближайшее время применительно к проблеме дипфейков на государственном уровне предлагается рассмотреть следующие предложения и практические шаги:

- создать целевую межведомственную экспертную группу по противодействию дезинформации и другим нарушениям с использованием дипфейков, что решило бы, в частности, задачу обмена передовым опытом внутри страны, в том числе по идентификации новых видов ДФ-инцидентов и защите от них;

- провести научно-обоснованную и практически полезную классификацию дипфейков по различным критериям, ввести единую терминологию и понятийный аппарат;

- сформировать и впоследствии обновлять специализированную базу данных по особо опасным видам дипфейков, в том числе с целью выявления характерных источников их наиболее массового распространения в российском сегменте Интернета (возможности создания такой базы сегодня рассматривают Центр глобальной ИТ-кооперации и Российская система качества);

- нацелить научно-исследовательские институты, другие профильные организации, венчурные фонды и стартапы на разработку методик и эффективных программных средств выявления ложного контента с целью сдерживания эпидемии дипфейков в наиболее значимых зонах цифровых медиа и в областях концентрации виртуальных социальных коммуникаций;

- внедрять всеми имеющимися законными средствами полезные практики и решения по дипфейкам на крупных национальных платформах и значимых информационных ресурсах, для чего разработать поэтапный план действий, скоординировать его с уже ведущейся работой по борьбе с деструктивным и противоправным контентом;

- опираться на существующие и развивать новые нормы саморегулирования, особенно в части эффективного модерирования контента на предмет дезинформации, подлога, клеветнических, мошеннических и других подобных материалов, а также контента, предназначенного для манипуляции и подрыва общественного порядка, что позволит частично сдержать негативные ДФ-процессы до выработки и принятия в стране соответствующих нормативно-правовых актов;

- при разработке новых законодательных инициатив по вопросам дипфей-

ков сохранять баланс между защитой прав человека, свободой слова и художественного самовыражения, с одной стороны, и мерами по защите частной жизни, общественного спокойствия и социального порядка, обеспечением комплекса вопросов безопасности на уровне государства и каждого гражданина — с другой;

- сформировать и довести до российских организаций, институтов и экспертов, осуществляющих международную работу, официальную государственную позицию и необходимые содержательные положения по вопросам дипфейков для соответствующего продвижения этих положений на международных площадках;

- предпринять в законодательной и правоприменительной практиках шаги по совершенствованию порядка принятия различных цифровых материалов в качестве доказательства в судах;

- урегулировать юридические вопросы, связанные с авторским правом применительно к различным видам дипфейков;

- **разработать руководства для жертв негативного ДФ-воздействия, в которые в том числе включить:**

1. информацию о возможных способах защиты прав и достоинства;
2. практические рекомендации с порядком конкретных действий по верификации сомнительной информации;
3. подробное описание процедуры обращения в официальные органы за помощью и юридической поддержкой;
4. другие полезные сведения и инструкции.

- рассмотреть в качестве совершенствования Меморандума о сотрудничестве в сфере охраны исключительных прав в эпоху развития цифровых технологий вопрос об обязательствах изымать из публичного доступа вредоносные дипфейки.

Приложение

Перечень публикаций для дополнительного изучения

В приложении приведены ключевые научные и аналитические публикации, в которых рассматриваются сложные взаимосвязи между важными понятиями в рамках рассматриваемой проблемы, а также разбираются философские, социальные, психологические и этические аспекты.

1. **Cristian Vaccari, Andrew Chadwick / «Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty and Trust in News» // Social Media // Society. — January–March 2020.** — https://journals.sagepub.com/doi/pdf/10.1177/2056305120903408.
2. **Haya R. Hasan, Khaled Salah / Combating Deepfake Videos Using Blockchain and Smart Contracts // Journal of Korea Multimedia Society. — August 2021.** — https://doi.org/10.9717/kmms.2021.24.8.1044.
3. **Ibrahim Mammadzada / «Deepfakes and Freedom of Expression: European Perspective» // Tallinn University of Technology, School of Business and Governance, Department of Law. — 2021.** — https://digikogu.taltech.ee/et/Download/ec6ea9ff-bd47-49e4-986d-3a115c00300b.
4. **Jan Kietzmann, Linda W. Lee, Ian P. McCarthy, Tim C. Kietzmann / «Deepfakes: Trick or Treat?» // University of Victoria // Nottingham Trent University // Simon Fraser University // Donders Institute for Brain, Cognition and Behaviour, Radbound University.** — https://core.ac.uk/download/pdf/250590695.pdf.
5. **Johannes Tammekänd, John Thomas, Kristjan Peterson / «Deepfakes 2020: The Tipping Point» // Sentinel. — October 2020.** — https://thesentinel.ai/media/Deepfakes%202020:%20The%20Tipping%20Point,%20Sentinel.pdf.
6. **J. Wojewidka / The Deepfake Threat to Face Biometrics // Biometric Technology Today. — November 2020.** — https://doi.org/10.1016/S0969-4765(20)30023-0.
7. **Katerina Kertysova / «Artificial Intelligence and Disinformation: How AI Changes the Way Disinformation Is Produced, Disseminated and Can Be Countered» // Security and Human Rights Monitor.** — https://www.shrmonitor.org/assets/uploads/2019/11/SHRM-Kertysova.pdf.
8. **Lyu Siwei / Detecting Deepfake Videos in The Blink of an Eye // The Conversation. — August 2018.** — http://theconversation.com/detecting-deepfake-videos-in-the-blink-of-an-eye-101072.
9. **Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, Yejin Choi / Defending Against Neural Fake News // Paul G. Allen School of Computer Science & Engineering // Allen Institute for Artificial Intelligence.** — https://proceedings.neurips.cc/paper/2019/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf.
10. **Max Read / «Can You Spot a Deepfake? Does It Matter?» // Intelligencer. — June 2019.** — http://nymag.com/intelligencer/2019/06/how-do-you-spot-a-deepfake-it-might-not-matter.html.
11. **Shadrack Awah Buo / «The Emerging Threats of Deepfake Attacks and Countermeasures» // Bournemouth University, Department of Computing and Informatics.** — https://arxiv.org/ftp/arxiv/papers/2012/2012.07989.pdf.
12. **Tom Simonite / «Will Deepfakes Disrupt the Midterm Election?» // Wired. — November 2018.** — https://www.wired.com/story/will-deepfakes-disrupt-the-midterm-election.
13. **Tim Hwang / «Deepfakes: A Grounded Threat Assessment» // Center for Security and Emerging Technology. — July 2020.** — https://cset.georgetown.edu/publication/deepfakes-a-grounded-threat-assessment/.

www.cgitc.ru

125009, г. Москва, Тверской бульвар, д. 14, стр. 1