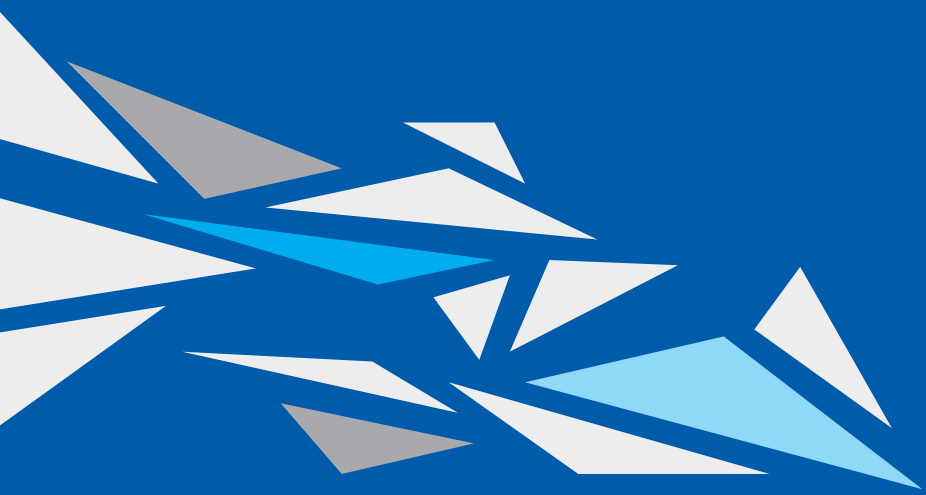
спринт фрии | Минцифры России

Дипфейки: риски и возможности для бизнеса



Москва 2024

спринт фрии | Минцифры России

Дипфейки: риски и возможности для бизнеса

Москва 2024

Содержание

Приветственное слово

Сергей Алимбеков
заместитель директора ФРИИ по технологическому развитию

Дорогие друзья!
Активное развитие новых коммуникационных интернет-технологий, технологий искусственного интеллекта, и массовое внедрение продуктов и сервисов на их основе имеет не только очевидную пользу, но и создает новые классы рисков и угроз, с которыми начинают сталкиваться как профессионалы, так и обычные люди – работающие, совершающие покупки, общающиеся и развлекающиеся в Интернете.

Развитие технологий для создания синтетического контента не только резко стимулировало развитие кинематографа, видеоигр, рекламы и различных видов цифрового творчества, но и дало недорогой и удобный инструмент различным мошенникам и аферистам. При этом простота и дешевизна этих инструментов позволяет применять их массово не только против «богатых и знаменитых», но и против самых обычных граждан и компаний малого и среднего бизнеса.

С другой стороны, если общество осознает риски дипфейков и готово тратить деньги на защиту от них, а технологические компании могут предложить решение данных проблем, то это открывает новую продуктовую нишу для специализированных продуктов и сервисов.

Для проверки этих гипотез мы обратились к профессиональным социологам из группы ЦИРКОН (АНО «Социологическая мастерская Задорина») и попросили их оценить готовность различных групп пользователей к противодействию злонамеренному применению дипфейков.
Результаты их исследований – в настоящей брошюре!

Введение

Эта брошюра подготовлена на основе отчета об аналитической работе по теме «Оценка готовности различных групп пользователей к применению специальных мер предотвращения рисков недобросовестного использования ИИ», выполненной АНО «Социологическая мастерская Задорина» в июле – сентябре 2024 года, а также анализа иностранных источников, выполненного командой ФРИИ.

Цель исследования – изучить возможности и перспективы возникновения и развития продуктовых ниш, связанных с обеспечением защиты массовых и профессиональных пользователей сервисов на основе НКИТ (социальных сетей и т. п.) от рисков, возникающих в результате недобросовестного и криминального использования технологий генерации или преобразования текстов, голоса, изображений и видео с помощью ИИ, на основе практик цифровой гигиены.

Исследование проводилось в два этапа.
Этап 1. *Кабинетное исследование.* Сбор и анализ данных из открытых источников о рисках, связанных с использованием дипфейков, а также о составе практик и навыков цифровой гигиены, снижающих риски от распространения дипфейков. В качестве источников данных были использованы отчеты KPMG и Центра глобальной ИТ-кооперации, статистика о потреблении дипфейков, исследования сервиса Sumsub и статьи экспертов в области ИИ-технологий, новостные материалы с информацией о примерах использования дипфейков в рекламе и иных сферах, статьи российских и зарубежных авторов об элементах цифровой гигиены, угрозах распространения дипфейков и др.

Этап 2. *Экспертная оценка.* В рамках опроса и серии интервью были собраны мнения 36 экспертов из пяти сфер:
1) представители ИТ-индустрии;
2) собственники, учредители (руководители) интернет-ресурсов;
3) представители онлайн-медиа, журналисты-обозреватели;
4) представители регулирующих/инвестиционных/управляющих структур;
5) представители исследовательского/академического сообществ.

Данное исследование направлено на экспертную проработку темы и не содержит утверждений о мнениях массовых аудиторий по исследуемому вопросу.

Термины и определения

В настоящем отчете применяют следующие термины с соответствующими определениями

- **Автоэнкодер –** технология (алгоритм), позволяющая создавать простые формы синтетического контента, такие как улучшение разрешения изображений и базовые манипуляции с текстовыми данными.
- **Дипфейк –** (от англ. deepfake – deep learning «глубинное обучение» + fake «подделка») – одна из наиболее заметных и обсуждаемых разновидностей так называемого синтетического контента – созданного с использованием глубоких нейронных сетей (DNN).
- **Живое присутствие** (от англ. facial liveness detection) – услуга по верификации биометрии.
- **Замена лиц** (от англ. faceswap – маска для лица) – метод, использующийся в создании дипфейков. Перенос черт лица одного человека на другого, оригинальные выражения лица и движения головы при этом сохраняются.
- **Реэнактмент** (от англ. reenactment – реконструкция) – метод, использующийся в создании дипфейков и позволяющий одному человеку контролировать мимику лица и движения головы другого, создавая эффект «кукольного мастера» (этот метод также называется сценарием puppet-master), когда внешний вид остается прежним, но управляется другим субъектом.
- **Фишинг** (от англ. fishing – рыбачить, выуживать) – вид социальной инженерии, в рамках которого злоумышленник осуществляет рассылки писем по электронной почте, в социальных сетях и мессенджерах пытается получить доступ к персональным данным пользователя, например, паролю от электронной почты или к информации о банковской карте.
- **Цифровая гигиена –** в широком смысле это набор навыков, помогающих избегать рисков, связанных с информационными технологиями.
- **FindFace –** потребительский сервис распознавания лиц, который использовал данные из социальной сети «ВКонтакте» и мог идентифицировать лица, сопоставляя их с 200 миллионами профилей в этой сети.
- **VisionLabs LUNA –** флагманский продукт, система распознавания лиц, которая интегрируется в системы безопасности банков и финансовых организаций, снижая риск мошенничества, связанного с подделкой личности при многофакторной аутентификации, когда помимо стандартных данных, например, паролей, для подтверждения личности используются биометрические данные, обработанные с применением дипфейков.

Перечень сокращений и обозначений

В настоящем отчете применяют следующие сокращения и обозначения

- **CISA –** Агентство по кибербезопасности и защите инфраструктуры
- **DNN –** глубокие нейронные сети (Deep Neural Network)
- **GANs –** концепция Generative Adversarial Networks
- **GPU –** мощные графические процессоры
- **KYC –** «Знай своего клиента» (Know Your Customer)
- **NIST –** Национальный институт стандартов и технологий США
- **NSA –** Национальное агентство безопасности США
- **ИИ –** искусственный интеллект
- **МТУСИ –** Московский технический университет связи и информатики
- **ПО –** программное обеспечение

Что такое дипфейки, кто и зачем их создает?

«Дипфейк как технология, как любой инструмент, любой топор, гвоздь, самолет, нож, пулемет и ядерный полураспад, всё это может быть использовано в мирных целях»

К. Р. Нигматуллина, доктор политических наук, профессор, заведующая кафедрой цифровых медиакоммуникаций СПбГУ

Дипфейк: к определению понятия

Дипфейк – калька с английского слова deepfake. Последнее составлено из двух частей: deep learning «глубокое обучение»[1], разновидность машинного обучения, предполагающая обучение искусственных нейронных сетей на больших объемах данных, и fake «фальшивка, подделка». В современный русский язык вошло слово «фейк», имеющее более узкое значение. Это не всякая фальшивка, но исключительно в цифровой информационной среде (близко по значению к англоязычному fake news).

Дипфейк представляет собой одну из наиболее заметных и обсуждаемых разновидностей так называемого синтетического контента – созданного с использованием глубоких нейронных сетей (DNN)[2]. Синтетический контент имеет широкий потенциал для использования, включая развлекательные и образовательные цели. Однако его бытование вызывает серьёзные опасения, связанные с нарушением этических норм, дезинформацией и киберпреступностью.

Считается, что слово «дипфейк» произошло от псевдонима пользователя популярного форума Reddit, в 2017 году опубликовавшего дипфейк-видео, которое привлекло внимание множества пользователей, в результате чего технология получила массовое распространение. При этом следует отметить, что технология использования машинного обучения в области компьютерного зрения была хорошо известна задолго до 2017 года – в киноиндустрии, индустрии видеоигр и научном сообществе. В 1997 году исследователи, работающие над технологией синхронизации губ, создали программу Video Rewrite, способную создавать видео с персонажем, «произносящим» текст, отличающийся от исходного. Хотя данная

[1] *DeepFake: как распознать и как защититься // Сбербанк. Режим доступа: www.sberbank.ru. Дата доступа: 30.09.2024.*
[2] *Mirsky, Y., & Lee, W. The creation and detection of deepfakes: A survey // ACM Computing Surveys (CSUR). 2021. Т. 54, № 1. С. 1-41.*

разработка была основана на машинном обучении, технология DNN в ней не использовалась, поэтому созданное таким образом видео в строгом смысле дипфейком не является.

«Давайте вообще начинать с того, что такое фейк и дипфейк, потому что цифровая гигиена так или иначе будет от этого отталкиваться. Как ни странно, выясняется, что никакого консенсуса в подходе к этим терминам ни у теоретиков, ни у практиков до сих пор нет; а между тем от их понимания зависит, с чем именно мы собираемся бороться... Тот же фейк распадается на несколько разновидностей – как минимум, есть понятия мисинформации (это ошибочная информация, но она связана с тем, что человек сам искренне заблуждается и в нее верит, он просто делится своим мнением, пусть и научно не до конца оправданным) и дезинформации (а это уже распространение ложных сведений с намеренной целью ввести в заблуждение, обмануть, здесь налицо злой умысел). И по поводу фейков в англоязычной литературе, а потом и в русскоязычной было очень много дискуссий: фейк – это мисинформация, или фейк – это только дезинформация, или фейк – это и то и другое? И примерно то же самое, как мне кажется, сейчас происходит с дипфейком: исследователи из разных стран спорят, называть ли дипфейком злостное создание недостоверного контента для того, чтобы получить какую-то выгоду и обмануть аудиторию (видео главным образом, но и аудио, и фото, и другие форматы тоже могут быть). Или же дипфейк – это просто наименование технологии, основанной на машинном обучении и подразумевающей аудио, видео или изображения людей, сгенерированные нейросетями?.. Например, когда известный актер продает право на создание своего цифрового клона для участия в рекламном ролике – технология есть, а обмана и злого умысла нет, – мы будем считать это дипфейком или нет? И это не просто отвлеченный академический вопрос. От этого зависят и решения крупных платформ по удалению недостоверного контента, и правовые кейсы, основанные на борьбе с ложной информацией... С моей точки зрения, дипфейк всё-таки – это определенная технология, а вот цели у дипфейка могут быть разными, в том числе довольно часто криминальные или манипулятивные».

С. А. Шомова, доктор политических наук, профессор Института медиа НИУ ВШЭ, исследователь Центра цифровых культур и медиаграмотности НИУ ВШЭ

Приведем несколько примеров определений дипфейка из российской научной и научно-практической литературы. В. Г. Иванов и Я. Р. Игнатовский понимают под дипфейком «методику компьютерного синтеза изображения, основанную на искусственном интеллекте, которая используется для соединения и наложения существующих изображений и видео на исходные изображения или видеоролики»[3].

По мнению А. С. Киселева, «дипфейками называют реалистичную замену лиц и голоса посредством использования генеративно-состязательных нейросетей»[4]. Н. Р. Красовская и А. А. Гуляев предлагают следующую формулировку: «совокупность технологических трансформаций изображения и видео, созданных с использованием искусственного интеллекта»[5].

Источник: vecteezy.com

[3] Иванов В. Г., Игнатовский Я. Р. Deepfakes: перспективы применения в политике и угрозы для личности и национальной безопасности // Вестник РУДН. Серия: Государственное и муниципальное управление. 2020. Т. 7. № 4. С. 379–386
[4] Киселев А.С. О необходимости правового регулирования искусственного интеллекта: дипфейк как угроза национальной безопасности // Вестник Московского государственного областного университета. Серия: Юриспруденция. 2021. № 3. С. 54–64.
[5] Красовская Н. Р., Гуляев А. А. Технологии манипуляции сознанием при использовании дипфейков как инструмента информационной войны в политической сфере // Власть. 2020. Т. 28. № 4. С. 93–98.

Приведенные определения весьма близки и не противоречат друг другу. Под дипфейком понимается либо метод обработки аудио и/или графики, предполагающий использование определенных технологий ИИ, либо медиаконтент, полученный с применением данного метода. Подчеркнем, что хотя слово «фейк» и в русском, и в английском языках несет негативную коннотацию, слово «дипфейк» характеризует определенный тип контента, вне зависимости от добросовестности или недобросовестности его применения.

Основные этапы развития синтетического контента

Развитие синтетического контента напрямую связано с прогрессом в области ИИ и машинного обучения. Рассмотрим основные этапы данного процесса.

Источник: vecteezy.com

Ранние эксперименты с синтетическим контентом

Раннее развитие дипфейков можно проследить до статьи 1997 года Кристофа Бреглера, Мишель Ковелл и Малкольма Слейни[6]. Этот документ заложил основу для разработки инновационной и уникальной программы, которая, по сути, автоматизировала работу киностудий. Предлагаемая программа создания видео могла синтезировать новую лицевую анимацию на основании анализа произносимых человеком фраз. Для изменения аудио и видео использовалось не только простое нелинейное редактирование видео (NLE), но и нейронные сети. Программа объединила предыдущую работу по интерпретации лиц, синтезированию звука из текста и моделированию губ в трехмерном пространстве.

В начале 2000-х исследователи внесли радикальные улучшения в создание реалистичных и убедительных дипфейков. Был предложен новый алгоритм под названием Active Appearance Models, который быстро завоевал популярность. Авторы алгоритма использовали статистическую модель для сопоставления формы с изображением, что значительно улучшило отслеживание черт лица. Такая модель опиралась на генеративно-состязательную сеть (GAN[7]) для выявления закономерностей на изображениях или видео и воссоздания лица цели в качестве результата.

В начале 2000-х исследователи внесли радикальные улучшения в создание реалистичных и убедительных дипфейков.

Благодаря быстрым улучшениям в этой области к 2016 году создание дипфейков можно было успешно осуществлять с использованием оборудования потребительского уровня. К этому моменту на рынке уже были разработаны методы синтеза человеческого голоса, и исследователи сосредоточились на манипулировании визуальными эффектами.

---

[6] Christopher Bregler, Michele Covell & Malcolm Slaney, Video Rewrite: Driving Visual Speech with Audio, ACM SIGGRAPH 1 (1997).

[7] GAN – алгоритм машинного обучения без учителя, построенный на комбинации из двух нейронных сетей, одна из которых генерирует образцы, а другая старается отличить правильные («подлинные») образцы от неправильных

Проекты Face2Face[8] и Synthesizing Obama[9] были добавлены к предыдущим вычислительным технологиям, чтобы обновить графическую точность и сделать видео реалистичным. Face2Face, проект Мюнхенского университета, создавал анимацию в реальном времени, заменяя область рта в целевом видео ртом актера.

Synthesizing Obama, проект команды Вашингтонского университета, ранее назывался Video Rewrite. Программа с улучшенной анимацией, текстурами и эмоциями. Графические улучшения затронули такие функции, как морщины и ямочки, они манипулировали цветами, чтобы лучше соответствовать освещению и оттенку кожи целевого видео. Алгоритмы были достаточно подробными.

Так, некоторые части лица, например брови, были очень точно синхронизированы с движением рта. Комбинация разработок позволила создать убедительный результат – модель с возможностью временного изменения как звука, так и видео. Оптимизация программ позволяла получить очень качественное видео длительностью в одну минуту в течение 40–50 минут на платформе с видеокартой NVIDIA TitanX и процессором Intel Core i7.

Появление Generative Adversarial Networks (GANs) (2014–2015 годы)

Значительный прорыв в создании синтетического контента произошел в 2014 году, когда студент Стэнфордского университета Ян Гудфеллоу и его коллеги предложили концепцию Generative Adversarial Networks (GANs)[10]. При использовании данного метода работают две нейросети.

Первая (генеративная, Generator, G) генерирует изображения, а вторая (дискриминативная, Discriminator, D) отвечает за поиск отличий между ними и исходными образцами. Эта технология быстро стала основой для создания реалистичных фейков, что привело к появлению новых технологий, таких как дипфейки.

[8] Face2Face: Real-time Face Capture and Reenactment of RGB Videos, PROC. CVPR (2016).
[9] Synthesizing Obama: Learning Lip Sync from Audio, 36(4) ACM TRANSACTIONS ON GRAPHICS 95:1 (2017).
[10] Goodfellow I., Pouget-Abadie J., Mirza M., et al. Generative adversarial nets // Advances in Neural Information Processing Systems. 2014. Т. 27. С. 2672–2680.

Замена лиц (faceswap) и реэнактмент (reenactment) – два ключевых метода создания дипфейков. Замена лиц предполагает перенос черт лица одного человека на другого, оригинальные выражения лица и движения головы при этом сохраняются. Реэнактмент позволяет одному человеку контролировать мимику лица и движения головы другого, создавая эффект «кукольного мастера» (этот метод также называется сценарием puppet-master), когда внешний вид остается прежним, но управляется другим субъектом. Внедрение GANs значительно улучшило качество дипфейков, особенно в области faceswap и reenactment. Ранним примером использования модели GAN был открытый инструмент pix2pix, который использовался для реэнактмента лиц.

Развитие синтетического аудиоконтента (2016–2017 годы)

Параллельно с развитием технологий генерации изображений и видео значительный прогресс был достигнут в области синтеза речи и аудио. В 2016 году Google представила WaveNet, модель глубокой нейронной сети, способную генерировать высококачественную синтезированную речь[11]. WaveNet смогла синтезировать речь, которая неотличима от настоящей, с точными интонациями и акцентами, что сделало возможным создание аудиодипфейков. Эта технология нашла применение в различных сферах, от голосовых помощников до автоматического озвучивания текстов.

Популяризация дипфейков (2017–2018 годы)

В период 2017–2018 годов технологии дипфейков стали массово известны благодаря созданию и распространению таких приложений, как FakeApp, которые позволяли пользователям легко создавать видео с заменой лиц. Это приложение получило широкую популярность и вызвало значительный общественный резонанс. В 2017 году, вскоре после своего запуска, FaceApp стало вирусным, число его загрузок превысило 80 миллионов по всему миру в течение первых месяцев. Только за июль 2019 года приложение было загружено свыше 12,7 млн раз в Google Play и App Store, став одним из самых загружаемых приложений того времени[12].

[11] Van Den Oord, A., Dieleman, S., Zen, H., et al. WaveNet: A generative model for raw audio. 2016. С. 1–15.
[12] Leskin P. Since going viral again for making people look old, FaceApp has been downloaded by 12.7 million new users // Business Insider. 2019. Режим доступа: www.businessinsider.com. Дата доступа: 23.09.2024.

Резкому увеличению числа дипфейков способствовал Reddit, запустив в 2017 году проект SubReddit «r/deepfakes», ныне удаленный. В нём во множестве появлялись порнографические дипфейки с участием многих известных актеров[13]. Лица знаменитостей на существующих порнографических видео добавляли с использованием алгоритма Face2Face, найденного в одной из библиотек с открытым исходным кодом. После закрытия r/deepfakes набирают популярность субреддиты непорнографического характера.

Расширение сфер применения и массовая интеграция (2019–2021 годы)

С 2019 по 2021 годы синтетический контент стал неотъемлемой частью цифровой среды. Одним из важнейших технологических достижений этого периода стало развитие и распространение языковых моделей, таких как GPT-2 и GPT-3 от OpenAI. В этот период продолжилось активное развитие технологий, связанных с созданием видеодипфейков. Модели DeepFaceLab и FaceSwap стали популярными инструментами для замены лиц и создания поддельных видео.

В коммерческой сфере одним из первых наглядных примеров использования технологии дипфейков стала разработка китайской компании Zao, которая давала возможность пользователям заменять лица на видеоизображениях знаменитостей или в сценах из популярных фильмов. Приложение быстро стало вирусным, но вызвало беспокойство по поводу конфиденциальности и безопасности данных, что привело к ограничению его использования в ряде стран.

Ярким примером применения дипфейков в этот период стало видео 2018 года с бывшим президентом США Бараком Обамой, созданное компанией BuzzFeed. Другой пример использования технологии дипфейков связан с выборной кампанией в США в 2020 году, когда дипфейки создавались для дезинформации и создания ложных новостей.

С годами доступ к технологии создания дипфейков стал проще. На GitHub есть множество библиотек и фреймворков, таких как TensorFlow, доступных для общественности, которые предоставляют

[13] *Samantha Cole, AI-Assisted Fake Porn Is Here and We're All Fucked, MOTHERBOARD (Dec. 11, 2017), www.vice.com.*

программное обеспечение для легкой разработки дипфейков. Приложения Zao, Faceswap, AvengeThem позволяют создавать ДФ на смартфонах.

Взрывной рост и современные вызовы (2022–2024 годы)

Всплеск фейков глобального масштаба пришелся на период пандемии. С 2022 года наблюдается экспоненциальный рост количества синтетического контента по всему миру. По данным компании Onfido, с 2022 по 2023 год число дипфейков увеличилось в 31 раз, что представляет собой рост на 3000%[14].

По данным компании Onfido, с 2022 по 2023 год число дипфейков увеличилось в 31 раз.

В значительной степени этот подъем обеспечен успехами технологических компаний в развитии генеративных нейросетей (генеративный ИИ), которые сделали системы на основе генеративного ИИ основным инструментом создания синтетических медиа. Подобные технологии теперь доступны в широко используемых потребительских инструментах, а не только в нишевых приложениях. Они просты в применении и, особенно в отношении аудио и изображений, не требуют навыков программирования, – могут быть проинструктированы (запрограммированы) простым языком. За 2023 год создание реалистичных изображений значительно улучшилось по качеству и настройке.

При этом имеющиеся на рынке коммерческие инструменты синтеза голоса позволяют не только генерировать абстрактные или заранее настроенные голоса, но и настраивать инструмент для генерации любого голоса на основе имеющихся записей. Так, с помощью широкодоступных инструментов клонирования звука достаточно одной минуты записи оригинального голоса, чтобы научить генеративную систему полностью имитировать этот голос.

[14] *Identity Fraud Report // Onfido. Режим доступа: www.onfido.com. Дата доступа: 23.09.2024.*

С помощью широкодоступных инструментов клонирования звука достаточно одной минуты записи оригинального голоса, чтобы научить генеративную систему полностью имитировать этот голос.

Для имитации видео возможности доступных коммерческих продуктов по-прежнему ограничены, им трудно добиться реалистичности в сложных реальных сценариях.

Вместе с тем общая тенденция развития генеративного ИИ указывает на всё возрастающую доступность для широкого круга пользователей простых инструментов для создания и вариаций реалистичных синтетических фотографий, аудио и видео конкретных реальных людей и контекстов.

Рисунок 1
Основные этапы развития синтетического контента

Социальные сети как среда распространения дипфейков

Массовому распространению дипфейков способствовало происходившее параллельно с развитием генеративных технологий совершенствование социальных сетей и встроенных в них рекомендательных алгоритмов.

Исследования показали, что фейковые новости распространяются в социальных сетях быстрее, чем достоверная информация. В одном из исследований MIT изучили 126 000 слухов, распространяемых тремя миллионами человек, и обнаружили, что ложные новости достигли большего числа людей, чем точная информация[15].

Это связано с тем, что шокирующие, необычные, скандальные новости вызывают больший интерес и желание поделиться у пользователей, чем привычная и традиционная информация. А так как выбор актуальных тем на онлайн-платформе обеспечивается алгоритмами, которые были настроены для сортировки, фильтрации и доставки контента таким образом, чтобы нарастить вовлечение пользователей, то в системе возникает положительная обратная связь: самое часто пересылаемое сообщение становится самым часто рекомендуемым к просмотру, а самое часто просматриваемое – часто пересылаемым.

В результате распространение информации растет в геометрической прогрессии независимо от того, правдива ли информация. **На онлайн-платформах отсутствует дополнительный фильтр для проверки источника и достоверности информации, как в традиционных источниках новостей, таких как СМИ.**

Другим аспектом, делающим социальные сети средой, удобной для распространения дипфейков, является их анонимность и возможность создания аккаунтов с любым именем, в том числе именами известных людей. Причём это касается не только текстовых социальных сетей, но и голосовых, в которых пользователи могут «узнавать» голоса известных людей и доверять им.

[15] Peter Dizikes, Study: On Twitter, false news travels faster than true stories, MIT NEWS (Mar. 8, 2018), www.perma.cc; Samantha Bradshaw & Phillip N.Howard, Why does Junk News Spread So Quickly Across Social Media, KNIGHT FOUNDATION (Jan. 29, 2018) www.perma.cc

С подобной проблемой в 2020 году столкнулся Clubhouse. Clubhouse предоставил платформу, на которой пользователи могли присоединяться к различным чатам по широкому кругу тем по своему выбору. Платформа была уникальна тем, что разговоры ведутся только в аудиоформате и исчезают навсегда, когда разговор в чате заканчивается.

Вскоре после запуска Clubhouse получил негативную реакцию из-за неспособности модерировать чаты должным образом. В январе 2021 года компанию обвинили в распространении на платформе теорий заговора о COVID-19[16]. Поскольку к каждой учетной записи привязан только голос, невозможно было доказать личность человека, стоящего за этим голосом.

Пользователи видят имя на странице профиля, связанной с голосом, и, если голос аккаунта похож на голос известного человека, то слушатели доверяют этому голосу в чате, даже если этот голос на самом деле этому человеку не принадлежит. Оказалось, что человеческое ухо очень легко обмануть. Реакция компании в виде изменения политики модерации последовала лишь через шесть месяцев.

Мобильные приложения, такие как Clubhouse, показывают, насколько легко обмануть пользователя и насколько сложно будет привлечь автора подмены к ответственности за свои действия на онлайн-платформах.

Clubhouse демонстрирует, что уровень анонимности, позволяющий использовать только свой голос, дает злоумышленникам больше возможностей воспользоваться аудиодипфейками. А учитывая запуск на волне ажиотажа вокруг Clubhouse аналогичных продуктов, предназначенных только для аудио, другими крупными социальными сетями, такими как Twitter, Facebook и Spotify, следует признать, что эта возможность применения аудиодипфейков стала доступна гораздо шире. Таким образом, нынешняя среда платформ социальных сетей позволяет дипфейкам развиваться и распространяться во всех их проявлениях.

[16] Yohance Kyles, Tiffany Haddish Responds To Accusations Of Bullying Black Doctor Over COVID-19 Conspiracy Theories, ALL HIPHOP (Jan. 15, 2021), www.perma.cc.

Области конструктивного использования дипфейков. Российский опыт

Если дипфейки так опасны, то возникает логичный вопрос: а почему бы их просто не запретить? В этом разделе мы собрали примеры полезного применения дипфейков.

Применение дипфейков разнообразно и может быть классифицировано по нескольким основным направлениям в соответствии с их функциями и последствиями использования. Наиболее активное применение дипфейков наблюдается в индустрии медиа и развлечений. Технология востребована для создания реалистичных симуляций знаменитостей, исторических личностей и вымышленных персонажей в рекламе, производстве фильмов и игровой индустрии.

Нестареющие и вечно живые звезды – это потенциальная возможность для рекламной индустрии. Сбер стал одним из первых российских брендов, использовавших дипфейк-технологию в рекламной кампании. При выборе актёра для рекламного ролика важным фактором становится его узнаваемость среди определенной целевой аудитории.

Так, в 2020 году компания создала рекламный ролик с участием персонажа Жоржа Милославского из фильма «Иван Васильевич меняет профессию». С помощью AI-маски и инструмента синтеза речи компания воссоздала образ популярного персонажа, который «рассказывал» о продуктах экосистемы бренда.
В 2021 году торговая сеть «Пятёрочка» применила технологию для создания ролика с участием актрисы Ольги Медынич.
Реклама стала успешной, её просмотрело 23 миллиона человек, что в 2 раза превысило прогнозируемые показатели.
По данным компании, использование дипфейков упростило производство, позволив сократить расходы на 20–60%[17].

В том же году «МегаФон» выпустил рекламный ролик тарифного плана «Первый семейный», в котором фигурирует образ голливудского актера Брюса Уиллиса.

[17] «Пятерочка» создала ролик с дипфейк-технологией для продвижения СТМ // AdIndex. ru. 22.12.2022. Режим доступа: www.adindex.ru. Дата доступа: 23.09.2024.

Для его создания специалисты загрузили в нейросеть 34 тыс. изображений Уиллиса, в том числе из «Крепкого орешка» и «Пятого элемента» с разных ракурсов для создания цифрового близнеца Уиллиса в высоком разрешении (4K). В рекламе также фигурирует телеведущий и юморист Азамат Мусагалиев. Ожидалось, что Брюс Уиллис привлечет более взрослую аудиторию старше 35 лет, а Мусагалиев заинтересует молодых людей.

В 2021 году торговая сеть «Пятёрочка» применила технологию для создания ролика с участием актрисы Ольги Медынич. Реклама стала успешной, её просмотрело 23 миллиона человек.

За создание цифровой версии Уиллиса отвечала российская компания Deepcake, созданная в 2020 году российскими инженерами, которой актер продал права на использование своего образа для создания цифровых двойников с помощью технологии генерации лиц. Это позволит артисту и в будущем появляться в новых проектах даже после того, как он принял решение уйти из кино из-за афазии.

По мнению некоторых экспертов, за разрешение на использование своей внешности Брюс Уиллис мог получить от «МегаФона» от одного до двух миллионов долларов[18].

В свою очередь, использование дипфейков может привести к снижению доверия потребителей к рекламным материалам. По данным исследования, проведенного Gallup в 2023 году, около 40% потребителей выразили недоверие к рекламе, в которой использовались дипфейки.

[18] Брюс Уиллис стал лицом «МегаФона» // Ведомости. 16.08.2021. Режим доступа: www.vedomosti.ru. Дата доступа: 23.09.2024.

Реакция властей, общества и ИТ-гигантов на проблему дипфейков

Рост числа дипфейков (ДФ) привел к необходимости разработки инструментов для их обнаружения. Sensity, основанная в 2018 году в Амстердаме, начала исследовать дипфейки и назвала себя «первой в мире компанией по визуальному анализу угроз». Тенденции показали, что количество доступных в Интернете дипфейков почти удваивается каждые шесть месяцев.

Несмотря на то, что первые дипфейки включали порнографический контент, последние популярные дипфейки нацелены на людей, которые были популярными политиками и интернет-знаменитостями. Комитет Палаты представителей по разведке США[19] в 2019 году провел открытые слушания по вопросу «угрозы национальной безопасности, создаваемой фейковым контентом с использованием искусственного интеллекта». Среди вопросов были методы обнаружения и борьбы с ДФ, оценка роли государственного и частного сектора, общества. DARPA в 2019 году финансирует проект медиакриминалистики[20], направленный на поиск способов автоматического выявления дипфейков.

Не остались в стороне ИТ-гиганты. Так, например, в 2020 году компания Microsoft запустила инструмент Video Authenticator, который способен анализировать видео и определять, является ли оно дипфейком, с помощью технологии анализа слоёв пикселей.

Социальные сети также начали интегрировать системы обнаружения дипфейков. В 2020 году Facebook и Instagram внедрили системы для автоматического обнаружения и маркировки фальшивых изображений и видео.

В сентябре 2019 года Facebook запустил конкурс Deepfake Detection Challenge (DFDC). Публичный конкурс поощрял людей к разработке автономных алгоритмических систем обнаружения дипфейковых видео. Участникам был предоставлен необработанный набор данных с видеозаписями за 38 дней, записанными 3500 актерами, некоторые из них были сфальсифицированными дипфейками.

[19] House Intelligence Committee on Intelligence
[20] govciomedia.com/darpa-launches-new-programs-to-detect-falsified-media

Более 2000 участников представили несколько моделей, каждая из которых имела новые алгоритмы обнаружения дипфейков. Победившая модель смогла обнаружить 82% дипфейков[21].

Полученная вероятность распознавания дипфейка в 82% может считаться обнадёживающей и показывающей, что техническое решение проблемы выявления синтетического контента может быть достигнуто в ближайшее время при концентрации на этой задаче значимого количества усилий и капиталов.

Однако стремительное развитие генеративных сервисов на основе ИИ поставило этот результат 2019 года под сомнение. Это косвенно подтверждается «перезагрузкой» множества государственных программ по борьбе с ДФ, в частности в США[22] в 2023 и 2024 годах, и почти полным исчезновением упоминаний о каких-либо достижениях в деле распознавания дипфейков. В том числе данных о достижении количественных результатов распознавания дипфейков, анонсированных в период до 2022 года различными государственными и окологосударственными агентствами.

Государственные стратегии борьбы с дипфейками и проблемы в их реализации

Гиперреализм дипфейков делает их идентификацию крайне трудной как для людей, так и для компьютерных технологий. Даже если дипфейк выявлен, найти и привлечь к ответственности виновного становится дополнительной проблемой, особенно с учетом отсутствия закона о дипфейках, в которых было бы прямо прописано, какие действия со сгенерированным контентом являются допустимыми, а какие нет, и кто должен нести ответственность за нарушение этих законов.

Проведенный анализ показывает, что деятельность правительств и корпораций по борьбе с дипфейками сводится к четырем направлениям:

[21] Deepfake Detection Challenge Dataset, FACEBOOK AI (June 25, 2020), www.perma.cc.
[22] DHS Launches First-of-its-Kind Initiative to Hire 50 Artificial Intelligence Experts in 2024, источник: DHS, февраль 2024, www.dhs.gov

Определение субъекта, ответственного за распространение дипфейков

В качестве такого субъекта американские законодатели определили социальные сети и коммуникационные платформы, возложив на них ответственность за выявление и удаление ДФ-контента, одновременно наделив их соответствующими полномочиями.

Это стимулировало технологические компании инвестировать в развитие технических средств выявления ДФ. В то же время это дало владельцам коммуникационных платформ основания для ничем не ограниченной цензуры в информационной среде под предлогом борьбы с ДФ.

В результате технологические гиганты начали внедрять политики и механизмы обнаружения дипфейков на своих платформах. У Meta[23], TikTok, Reddit и YouTube есть политика модерации контента, которая призывает к запрету и удалению синтетических медиа, создающих вводящий в заблуждение контент (хотя президентская кампания 2024 года на сегодня выявила недостатки в защите).

Meta[23] разработала ИИ-инструмент, который занимается обратным проектированием изображений, созданных искусственным интеллектом, для отслеживания их происхождения. Google выпустила большой набор данных визуальных дипфейков (включенный в тест FaceForensics – набор данных для обнаружения дипфейков). Microsoft запустила Microsoft Video Authenticator, который обеспечивает процент вероятности того, что фото или видео подверглись искусственной манипуляции.

В сентябре 2023 года Google также объявила, что проверенные рекламодатели, публикующие рекламу к выборам, должны будут помечать искусственно обработанный контент. Некоторые компании предприняли усилия по стимулированию разработки технологий обнаружения дипфейков, таких как технология Meta Deep Fake Detection Challenge, в то время как другие, такие как Google, запретили обучение систем искусственного интеллекта созданию дипфейков.

[23] Деятельность запрещена на территории России.

В то же время европейские законодатели обозначили в качестве ответственного за криминальное использование дипфейка только того, кто его применил в этом качестве. А не саму сеть или коммуникационную платформу.

Такой подход[24] в большей степени соответствует принципам свободы слова, но переносит бремя проверки подлинности информации на пользователя.

Государственные инвестиции в развитие технологий обнаружения ДФ-контента

В США положение Закона о полномочиях национальной обороны (NDAA) 2020 года предусматривает, что директор национальной разведки должен подготовить подробный отчет об использовании дипфейков в качестве оружия.

Закон об определении результатов генеративно-состязательных сетей 2020 года поручил Национальному научному фонду (NSF) и Национальному институту стандартов и технологий (NIST) поддержать исследования в области GAN, а Агентство перспективных исследовательских проектов Министерства обороны (DARPA) создало две программы: MediFor (заключение в 2021 финансовом году) и SemaFor (в настоящее время), посвящённые обнаружению дипфейков[25]. По состоянию на 2024 год результаты работы этих программ в публичном информационном поле не отражены.

Введение на законодательном уровне обязательной маркировки ДФ-контента

Данный подход строится на предположении, что в ядре многочисленных сервисов и программных продуктов, создающих дипфейки, находится относительно небольшое число генеративных моделей, принадлежащих ограниченному количеству разработчиков, большинство из которых являются американскими компаниями или действующими в американской юрисдикции. Обязав их маркировать дипфейки и вообще синтетический контент[26], американские

[24] Artificial Intelligence Act: MEPs adopt landmark law, Europen Parlament News, мрат 2024.
[25] Kelley M. Sayler and Laurie A. Harris, Deep Fakes and National Security, Congressional Research Service , April 17, 2023.
[26] H.R.5586 - DEEPFAKES Accountability Act, 118th Congress (2023–2024), USA. Источник – www.congress.gov

законодатели надеются резко сократить количество случаев нежелательного контента[27].

Естественным ограничением данного подхода является появление всё большего количества дипфейк-сервисов в странах, не признающих американское законодательство, и целенаправленно разрабатывающих дипфейки как инструменты информационного противостояния с США. Аналогично можно ожидать, что дипфейки, создаваемые в рамках информационных войн, проводимых в интересах США, также не будут иметь соответствующей маркировки.

Просвещение населения о возможностях ДФ

Американские законодатели рекомендуют правительству и некоммерческим организациям приложить усилия по повышению цифровой грамотности населения[28, 29], чтобы информировать общественность о явных признаках дипфейков, а также о практике поиска вторичных и третичных источников для подтверждения информации или контента. Такие «предварительные» меры доказали свою эффективность.

Подобные меры, повышая устойчивость населения к ДФ, одновременно с этим снижают уровень его доверия к любой распространяемой информации.

Этот подход, основанный на повышении навыков населения, необходимых для безопасной жизни в цифровой среде – цифровая гигиена, – также поддерживается российскими экспертами и специалистами по кибербезопасности.

В частности, подробной разработке этого подхода посвящена работа Ашманова И. С., Касперской Н. И. «Цифровая гигиена», изданная в 2021 году. Более подробно возможности и ограничения цифровой гигиены рассмотрены в следующий главе.

[27] USA – 11 States Now Have Laws Limiting Artificial Intelligence, Deep Fakes, and Synthetic Media in Political Advertising – Looking at the Issues, источник – www.lexology.com.
[28] Quality Media Literacy Requires More Than Toothless Laws, 2024, источник: www.edutopia.org.
[29] Defending Against Deep Fakes Through Technological Detection, Media Literacy, and Laws and Regulations, источник: THE INTERNATIONAL AFFAIRS REVIEW - www.iar-gwu.org. Источник – www.congress.gov

Практики и навыки цифровой гигиены пользователей, снижающие риски от дипфейков

Цифровая гигиена: определение

С появлением Интернета и развитием цифровых технологий, включая технологии искусственного интеллекта, вопросы сохранения персональных данных, защиты от мошенничества и сохранения психического здоровья становятся всё актуальнее. По данным МВД России за 2020 год, число преступлений, совершенных с использованием информационно-телекоммуникационных технологий, выросло на 73,4%.

Общее количество таких зарегистрированных преступлений – 510 396, значительную их часть составляют мошенничество – 237 074 – и кражи – 173 416[30]. Как отмечает источник, уровень раскрытия таких преступлений остается весьма низким, так как у ведомства не хватает необходимых ресурсов и специалистов. При этом меры предотвращения и защиты от подобных угроз обсуждаются на самом высоком уровне.

Стратегия национальной безопасности Российской Федерации (утверждена Указом Президента РФ № 400 от 2 июля 2021 г.) подчеркивает важность создания безопасного информационного пространства и защиты российского общества от негативного информационно-психологического воздействия как одного из ключевых национальных интересов России[31]. В начале 2021 года Президент РФ поручил разработать концепцию защиты прав и свобод человека в цифровом пространстве и план мероприятий по ее реализации, включая повышение цифровой грамотности и обучение информационной безопасности[32]. Специалисты в области кибербезопасности также отмечают цифровую гигиену как основной способ предупреждения и борьбы с похищением данных и мошенничеством, поскольку собственная внимательность и осведомленность является лучшим способом защиты пользователя. Цифровая гигиена в широком смысле – это набор навыков, помогающих избегать рисков, связанных с информационными

[30] Цифровая безопасность личности: что изменилось за последний год // Гарант. Режим доступа: www.garant.ru. Дата доступа: 23.09.2024.
[31] Указ Президента РФ от 2 июля 2021 г. N 400 «О Стратегии национальной безопасности Российской Федерации» // Гарант. Режим доступа: www.base.garant.ru. Дата доступа: 23.09.2024.
[32] Поручение Президента РФ от 28 января 2021 г. // Гарант. Режим доступа: www.base.garant.ru. Дата доступа: 23.09.2024.

технологиями[33]. В узком смысле она связана с информационной безопасностью и защитой от цифрового мошенничества. С научно-медицинской точки зрения, цифровая гигиена включает разработку стандартов и мер для улучшения информационной среды, чтобы уменьшить негативное влияние технологий на физическое и психическое здоровье и благополучие людей и общества. Понятие цифровой гигиены появилось вместе с развитием цифровых ресурсов (телефоны, смартфоны, Интернет). По аналогии с общепринятой гигиеной цифровая призвана сохранить здоровье и благополучие человека в цифровом мире.

«Для меня цифровая гигиена, если продолжать медицинские метафоры, – это навык санировать собственное цифровое и информационное пространство. Это умение проверять ту информацию, которая кажется тебе не совсем убедительной, и способность отличать то, что непременно нужно верифицировать, от информации «проходной» и/или достаточно достоверной. При этом, как вы понимаете, проверять всё подряд невозможно (я не могу себе представить такого человека, который будет проверять абсолютно всё, с чем он сталкивается в инфополе). Видимо, мы проверяем то, что для нас очень актуально, то, что задевает нас лично, то, что связано с нашим каким-то человеческим интересом. Получается, что цифровая гигиена – это всегда самоограничение, это умение выбирать, умение отличать ту информацию, которую можно просто пробежать глазами, от той информации, которую надо обязательно проверить, а также знание инструментов верификации контента».

С. А. Шомова, доктор политических наук, профессор Института медиа НИУ ВШЭ, исследователь Центра цифровых культур и медиаграмотности НИУ ВШЭ

Вместе с тем эксперты отмечают и ряд других важных аспектов. Цифровую гигиену стоит рассматривать не только как средство защиты от угроз онлайн-среды, но и как набор мероприятий по сохранению активной позиции в цифровом пространстве.

[33] Суходаева Т.С. Формирование навыков цифровой гигиены студентов как существенный элемент воспитательной работы в высшей школе // Вестник СГУПС: гуманитарные исследования. 2022. №2 (13). С. 100-105.

При оценивании изменения практик цифровой гигиены на фоне распространения дипфейков мнения экспертов разделились.

Одни считают, что активное распространение дипфейков и появление новых цифровых угроз требует пересмотра прикладных правил цифровой гигиены (при сохранении общих правил «сохранять здравый смысл, относиться к источнику с недоверием»).

Вторые, напротив, говорят о том, что уровень угрозы, которую несут дипфейки, ещё не настолько критичен, чтобы кардинально менять систему практик цифровой гигиены.

Третьи выражают более скептическую позицию и говорят о том, что в краткосрочной перспективе рост числа синтетического контента фактически нивелирует защитные возможности использования правил цифровой гигиены и потребует новых инфраструктурных решений от государства и бизнеса.

Описание состава навыков цифровой гигиены пользователей

Меры цифровой гигиены направлены на предотвращение и профилактику угроз, возникающих в процессе использования электронных ресурсов (киберугрозы). К основным угрозам, с которыми сталкиваются пользователи в Интернете, можно отнести кражу личных данных и информации третьими лицами; мошенничество; потерю психологического благополучия и ментального здоровья. Базовые правила цифровой гигиены направлены на предотвращение утечки данных, защиту от мошенничества и сохранение психологического благополучия. К ним относятся: быть осторожным с личной информацией; использовать надежные пароли и двухфакторную аутентификацию; быть внимательным к незнакомцам; проверять источники информации и доверять только проверенным; обращать внимание на время, проведенное в сети[34, 35]. Эти меры – универсальны, однако степень уязвимости пользователей отличается в зависимости

[34] Цифровая гигиена // Российская газета. 05.01.2024. Режим доступа: www.rg.ru. Дата доступа: 23.09.2024.
[35] Ашманов И.С., Касперская Н.И. Цифровая гигиена. СПб: Издательский дом «Питер», 2021.

от возраста, образования, видов и частоты использования цифровых ресурсов. Среди наиболее уязвимых категорий отмечаются дети и лица пожилого возраста[36].

С развитием технологий способы мошенничества и обмана в цифровом пространстве становятся всё более изощренными. Создаются поддельные сайты, фальшивые новости, организации и люди. В этой связи базовых мер цифрой гигиены становится недостаточно для борьбы с манипуляциями и фальшивой информацией (фейки и дипфейки).

В контексте защиты от дипфейков рекомендуется относиться критически ко всей аудио- и видеоинформации и использовать дополнительные источники проверки данных.

Можно выделить следующие базовые правила защиты от дипфейков. Во-первых, нужно знать о существовании таких технологических манипуляций. Во-вторых, необходимо быть внимательными при просмотре видео и аудиозаписей в Интернете.

Несмотря на высокую реалистичность, дипфейки могут содержать небольшие артефакты, такие как неестественные движения глаз, странные выражения лица или несоответствия в освещении, которые могут выдать подделку.

В-третьих, необходимо проверять источники информации (надежность и репутация источника), чтобы убедиться, что представленная информация является достоверной. В контексте защиты от дипфейков рекомендуется относиться критически ко всей аудио- и видеоинформации и использовать дополнительные источники проверки данных (официальные новостные источники, звонок на официальный номер организации или личный номер человека и его родственников).

[36] *Good Digital Hygiene Practices You Need to Know // Vodafone. Режим доступа: www.vodafone.com. Дата доступа: 23.09.2024.*

Меры цифровой гигиены для защиты от дипфейков различаются в подходах и акцентах. Некоторые источники делают упор на использование передовых технологий для выявления дипфейков[37, 38], другие источники[39, 40] больше фокусируются на базовых принципах цифровой гигиены, не углубляясь в технические детали (мыслить критически и понизить уровень доверия к источникам из Интернета).

К технологическим способам защиты от дипфейков относятся меры, задействующие использование какого-либо программного обеспечения (ПО) для защиты учетных записей или устройства.

К ним относятся: использование сильных паролей и двухфакторной аутентификации; регулярное обновление программного обеспечения и установка антивирусов; использование отдельных учетных записей для разных целей.

Если говорить о специфических мерах по противодействию синтетическому контенту, то в настоящий момент основное технологическое средство защиты – это программы по его распознаванию.

Существующие на рынке программы по распознаванию синтетического контента, по оценкам экспертов, способны определять его с вероятностью от 28 до 86%.

Отмечая отдельные программы для противодействия дипфейкам, эксперты отмечают не только сравнительно невысокую эффективность существующих технических средств защиты, в частности, по обнаружению синтетического контента, но и в целом отсутствие на рынке известных и популярных продуктов, которые могли бы обеспечить защиту пользователя от недобросовестного использования ИИ.

[37] *Good Digital Hygiene Practices You Need to Know // Vodafone. Режим доступа: www.vodafone.com. Дата доступа: 23.09.2024.*
[38] *Semantic Forensics // DARPA. Режим доступа: www.darpa.mil. Дата доступа: 23.09.2024.*
[39] *Dartmouth Guide to Digital Hygiene // Dartmouth College. Режим доступа: www.services.dartmouth.edu. Дата доступа: 23.09.2024.*
[40] *Digital hygiene // Group-IB. Режим доступа: www.group-ib.medium.com. Дата доступа: 23.09.2024.*

«В определенной мере мы оказались беззащитны перед лицом дипфейков и более изощренных методов обмана. Человеческому восприятию становится всё сложнее отличить подлинное от искусственно созданного. Будь то телефонные звонки с предложением услуг, записанные видео в мессенджерах или попытки дестабилизировать политическую обстановку – все эти инструменты теперь могут быть использованы злоумышленниками с невиданной ранее достоверностью. Современные алгоритмы искусственного интеллекта так хороши в генерации контента, что даже специалисты по кибербезопасности не успевают реагировать на новые угрозы. Мы стоим на пороге эпохи, где грань между реальностью и искусной имитацией становится всё более размытой».

Д. А. Касьяненко, старший преподаватель, эксперт факультета компьютерных наук НИУ ВШЭ

В этой связи единственной возможной альтернативной защитой на данный момент являются поведенческие рекомендации, которые направлены на помощь пользователю в распознавании ложной информации и предотвращении обмана на индивидуальном уровне:

- защищать личную информацию и проявлять осторожность при предоставлении личной информации онлайн (не делиться критической информацией, использовать псевдонимы, избегать перехода по подозрительным ссылкам);
- загружать приложения и файлы только из официальных источников;
- избегать хранения важной личной информации и документов (например, фото паспорта) в мессенджерах, социальных сетях;
- не перезванивать по предоставленным номерам, звонить только по номерам, известным заранее и из надежных источников (например, звонить самому только по публично известным номерам службы поддержки банков)[41, 42];
- мониторить и отслеживать упоминания о себе в социальных сетях и Интернете вообще;

[41] *Digital hygiene: the most important unfinished business // Telefonica. Режим доступа: www.telefonica.com. Дата доступа: 23.09.2024.*
[42] *Good Digital Hygiene Practices You Need to Know // Vodafone. Режим доступа: www.vodafone.com. Дата доступа: 23.09.2024.*

- проверять информацию, полученную из онлайн-каналов через другие источники (например, посредством телефонного звонка, видеозвонка или очной встречи);
- проявлять осторожность к голосовым сообщениям. Следует быть внимательным к подозрительным голосовым сообщениям, особенно если они приходят от известных номеров[43]. Так как мошенники могут взламывать профили в мессенджерах, а затем отправлять фейковые голосовые сообщения контактам взломанного человека;
- получить собственный опыт создания и анализа синтетического контента.

«На одном из наших очных занятий мы поставили перед студентами задачу: используя HeyGen, платформу для создания цифровых двойников, незаметно снять видео другого студента, в котором его «двойник» просит 10 тысяч рублей. Все успешно справились с заданием, и участники с удивлением смотрели на результат: «это же я». Это позволяет нам воссоздать путь злоумышленника и убедиться в том, насколько эти технологии действительно эффективно имитируют реальность».

Д. А. Касьяненко, старший преподаватель, эксперт факультета компьютерных наук НИУ ВШЭ

Обобщая сказанное, можно выделить следующие базовые рекомендации для индивидуальных пользователей, компаний и регулирующих органов (таблица 1).
Пункты в таблице не ранжируются по важности. На индивидуальном уровне особое внимание уделяется навыкам недоверия незнакомым источникам и защите своих личных данных.

Человеческому восприятию становится всё сложнее отличить подлинное от искусственно созданного.

[43] *Digital hygiene: the most important unfinished business // Telefonica. Режим доступа: www.telefonica.com. Дата доступа: 23.09.2024.*

Таблица 1

Основные правила цифровой гигиены[44]

№	Для индивида	Для организации	Для регулирующих органов
1	*Критическое отношение ко всей информации от незнакомых источников*	Внедрение политики использования личных устройств	*Разработка законодательства, регулирующего использование ИИ-технологий (в том числе маркировка контента, созданного ИИ)*
2	Безопасность паролей: использование уникальных паролей и их регулярная смена	*Регулярное обучение и информирование сотрудников о типах угроз и способах защиты*	*Стимулирование сотрудничества между ведомствами, частными и государственными компаниями для борьбы с фейковой информацией и киберугрозами*
3	*Защита личных данных: не выкладывать конфиденциальную информацию даже в закрытые источники*	*Использование систем безопасности: программные или аппаратные решения, защищающие сеть организации от внешних угроз*	*Образовательные инициативы для повышения осведомленности населения*
4	*Осторожность в сети: открывать только проверенные ссылки, скачивать приложения из надежных источников*	Защита данных: создание копий важных данных и их шифрование, использование надежного ПО	*Поддержка исследований и разработок в создании технологических средств защиты от киберугроз*
5	Цифровое благополучие: контроль времени, проведенного в сети, и просматриваемого контента	Работа с третьими сторонами: проверка систем безопасности партнеров и заключение договоров о конфиденциальности	

[44] Курсивом выделены правила цифровой гигиены, применимые для защиты от рисков и угроз, вызванных использованием дипфейков.

На корпоративном уровне требуется внедрение системы безопасности, которая включает в себя как технологические инструменты распознавания и защиты от дипфейков, так и развитие навыков цифровой гигиены сотрудников. Основой развития навыков цифровой гигиены является обучение людей и повышение уровня их осведомленности о существующих технологиях и возможных угрозах. Обучение может включать как отдельные мастер-классы для детей и взрослых, так и специальные курсы для студентов и сотрудников отдельных организаций.

Кроме того, для ряда компаний тенденция к отказу от использования зарубежных ресурсов является актуальным аспектом цифровой безопасности. Например, отказ от использования Zoom для проведения видеоконференций в пользу отечественных аналогов, переход с Google Docs на локальные системы документооборота, использование национальных социальных сетей вместо зарубежных платформ[45].

Причины таких решений связаны с опасениями, что зарубежные сервисы могут быть использованы для шпионажа или сбора конфиденциальной информации. Помимо этого, законодательство некоторых стран может требовать хранения данных на территории их страны.

Трудности, которые возникают при следовании правилам цифровой гигиены

Цифровая гигиена – важная профилактическая мера для защиты от киберугроз на разных уровнях, однако эффективность представленных рекомендаций имеет ряд ограничений.

Во-первых, зачастую рекомендации носят общий характер (не открывать подозрительные ссылки, не доверять недостоверным источникам). Рекомендации не учитывают конкретную ситуацию или организацию, поэтому пользователь зачастую может не ассоциировать свою ситуацию с критической.

[45] Российских учителей и чиновников переведут на отечественные мессенджеры. Режим доступа: www.cnews.ru. Дата доступа: 23.09.2024.

Во-вторых, некоторые рекомендации могут противоречить друг другу или быть несовместимыми с определенными рабочими процессами. Например, не отвечать на звонки с незнакомых номеров или не перезванивать на незнакомые номера. Некоторые рекомендации являются трудновыполнимыми. Например, использовать сложные пароли для каждого аккаунта и не хранить их в электронном виде.

Некоторые рекомендации являются сложными для понимания отдельными пользователями или нереализуемыми ввиду высокой стоимости.

В-третьих, некоторые рекомендации являются сложными для понимания отдельными пользователями или нереализуемыми ввиду высокой стоимости. Например, настройка двухфакторной аутентификации, шифрование данных и установка комплексных систем безопасности. Излишний алармизм в некоторых рекомендациях по цифровой гигиене отмечают и эксперты, предлагая искать компромисс между удобством и безопасностью для пользователя:

«У нас всегда есть компромисс между удобством и безопасностью, и в этом смысле все инструменты должны быть устроены так, чтобы они максимизировали безопасность, но минимизировали дискомфорт, который связан с исполнением мероприятий по цифровой гигиене».

Д. В. Сошников, кандидат физико-математических наук, доцент факультета компьютерных наук НИУ ВШЭ, доцент МАИ

Тем не менее подобные рекомендации важны для обеспечения базовой защиты и повышения общей осведомлённости. Эффективность рекомендаций может быть выше, если они будут более конкретными, сформулированными понятным языком, и станут предметом регулярного обучения пользователей, а также, если появятся цифровые сервисы, которые помогут эти рекомендации соблюдать.

Механизмы формирования у целевых групп навыков цифровой гигиены

Особо уязвимые группы населения

Среди уязвимых целевых групп эксперты, как правило, отмечают детей, подростков и пожилых людей. Первые фактически выросли в цифровой среде, однако пока не обладают в достаточной степени социальными навыками, необходимыми, чтобы обезопасить себя в цифровом пространстве. Вторые, напротив, не обладают высоким уровнем цифровых навыков, и тоже нуждаются в формировании новых социальных компетенций.

Данные целевые группы требуют специальных мер и механизмов распространения и освоения навыков цифровой гигиены. Механизмы формирования навыков цифровой гигиены соответствуют механизмам формирования цифровой грамотности населения в целом и отдельных целевых групп. Речь идет о механизмах социальной адаптации, включающих передачу социально полезной информации «из уст в уста», а также через медийные каналы коммуникации.

Во втором случае особая роль принадлежит профессиональным коммуникаторам – журналистам, а также лидерам мнений – персонам, обладающим экспертным статусом или претендующим на него: ИТ-специалистам, блогерам. В данном контексте предполагается появление суперэкспертов – медийных персон, активно высказывающихся по вопросам цифровой гигиены. В настоящее время соответствующая позиция в информационном поле вакантна, ближе всего к ней находится И. С. Ашманов. Поддержка информационно-коммуникационных программ повышения навыков цифровой гигиены может исходить от ИТ-бизнеса, заинтересованного в повышении безопасности использования предлагаемых цифровых продуктов.

Для отдельных целевых аудиторий возможно действие специфических механизмов формирования навыков цифровой гигиены.
В корпоративной среде это члены трудовых коллективов (коммуникации peer-to-peer), а также специализированные подразделения, отвечающие за корпоративную культуру и безопасность.
Для подростковой и молодежной среды наряду с родителями важную роль играют педагоги, представители администрации образовательных

учреждений и другие значимые взрослые. Для представителей старших возрастных групп – референтные представители младшего поколения из близкого окружения и статусные медийные эксперты.

Технологические решения, которые помогут в реализации правил цифровой гигиены

Представления экспертов о возможных технологических решениях, которые помогали бы соблюдать цифровую гигиену, ещё достаточно абстрактные, некоторые из них существуют только на уровне идеи, не прототипа. Тем не менее уже можно выделить ряд технологических решений, которые помогут в соблюдении рекомендаций по цифровой гигиене индивидуальным пользователям.

Таблица 2

Возможные технологические решения для соблюдения правил цифровой гигиены

№	Рекомендация для индивида	Технологическое решение
1	Критическое отношение ко всей информации от незнакомых источников	• Персонализированные ассистенты (индивидуальные трекеры рисков/угроз), которые маркируют потенциально опасный контент • Образовательные продукты по обучению использования ИИ
2	Безопасность паролей: использование уникальных паролей и их регулярная смена	• Приоритизация хранения паролей (пристальное обеспечение особой безопасности только тех паролей, утеря которых может понести наибольший ущерб) • One Time Password Generator (OTP), генераторы одноразовых паролей
3	Защита личных данных: не выкладывать конфиденциальную информацию даже в закрытые источники	• Антивирусные программы • ИИ-сканер разговоров и маркирование злоумышленников (без сохранения содержимого звонка) • Антифрод-системы на стороне мобильных операторов
4	Осторожность в сети: открывать только проверенные ссылки, скачивать приложения из надежных источников	• Программы, которые маркируют подозрительные ссылки (спам-фильтр) • Маркировка контента, произведенного физическим лицом (например, с помощью ЭЦП)
5	Цифровое благополучие: контроль времени, проведенного в сети, и просматриваемого контента	• «Приложения-скафандры», которые будут замещать человека в онлайн-взаимодействиях

Стоит отметить, что часть разработок, названных экспертами, основывается на технологиях ИИ. Неоднократно эксперты упоминали персонализированных ассистентов на основе ИИ, которым можно делегировать часть задач по проверке контента:

«Можно представить себе такого искусственного агента, который читает нашу почту, нашу переписку, собирает все сообщения вместе и какую-то уже персонализированную рекомендацию выдает».

Д. В. Сошников, кандидат физико-математических наук, доцент факультета компьютерных наук НИУ ВШЭ, доцент МАИ

Примечательно, что эксперты не упоминали в качестве возможного эффективного технологического решения маркировку контента, созданного ИИ, которую довольно часто обсуждают в публичном пространстве. Один из экспертов высказал мнение, что возможным решением в борьбе с недобросовестным контентом может стать маркировка контента, создаваемая не ИИ, а автором – физическим лицом, вероятно, с использованием ЭЦП.

Источник: freepik.com

«Например, я стою на позиции, что единственный способ борьбы с дипфейками – это публикация или использование того контента, который подписан автором. То есть, если есть кому потом предъявить претензии за то, что там была написана ложь, или картинка не такая, какой должна быть, но человек при публикации этой картинки или текста поставил свою подпись, и сказал, что я ручаюсь, что это утверждение правильное, то с таким контентом можно иметь дело. Есть, конечно, всякие такие хитрые псевдоинструменты, которые позволяют отличить контент фейковый от нефейкового.

Единственный способ борьбы с дипфейками – это публикация или использование того контента, который подписан автором.

Там достаточно ограниченный набор, используется для этого различения тоже искусственный интеллект, и оттого использование этих мер кажется мне достаточным тупиком, потому что борьба искусственного интеллекта против искусственного интеллекта безнадежна. Никаких надежных мер защиты от искусственного интеллекта и оценки доверенности искусственного интеллекта пока не существует... С помощью ЭЦП можно картинку подписать, можно текст подписать, можно аудио подписать, всё же это бинарные файлы. Документ – такой же бинарный файл. Просто вы привыкли к тому, что ЭЦП применяется во всяких системах электронного документооборота, а в этих системах электронного документооборота есть еще дополнительные средства для отображения тех документов, которые мы подписываем. Но с тем же успехом можно сделать систему для отображения картинок, воспроизведения звуков подписанных»

С. В. Белова, генеральный директор IDX

Однако делегирование функции защиты от недобросовестного контента ИИ тоже сопряжено с рисками, например, потери контроля над алгоритмами, распознающими и фильтрующими контент.

«Вы знаете, если бы вы меня спросили лет 5–8 назад, когда я очень верила в медиаграмотность как первооснову для понимания и анализа медиа, я бы сказала, что, безусловно, важнее всего просвещение. В первую очередь надо учить критическому мышлению, и, собственно, мы этим и занимались все эти годы. Но сегодня я должна сказать, что чем изощреннее становятся технологии и чем больше я понимаю, что человеческий мозг просто так устроен, что он ведется на определенного рода манипуляции, – тем больше я прихожу к выводу, что медиаграмотность – это всё-таки не панацея, ее возможности ограничены».

С. А. Шомова, доктор политических наук, профессор Института медиа НИУ ВШЭ, исследователь Центра цифровых культур и медиаграмотности НИУ ВШЭ

Помимо технологических решений, которые называли эксперты, можно отметить ещё ряд способов защиты от недобросовестного использования дипфейков, которые не требуют дополнительных разработок.

Например, при получении запроса на перевод денежных средств от знакомого лица всегда стоит запрашивать конфиденциальную информацию (которую знает только «настоящий» коммуникант) или искать выхода на реальный (физический) контакт. А при столкновении с шантажом эксперты рекомендуют публиковать сгенерированные дипфейки первым («не вовлекаться») и не отвечать на сообщения шантажистов.

В целом, можно отметить, что из-за отсутствия в настоящее время эффективных технологических решений для противодействия дипфейкам эксперты отдают преимущество рекомендациям, практика следования которым не требует цифровых посредников. Однако отмечают, что следовать таким рекомендациям по мере совершенствования технологий для использования в недобросовестных целях становится всё сложнее.

Интегральная оценка возможности снижения рисков от использования дипфейков и оценка готовности различных групп пользователей к применению специальных мер предотвращения рисков недобросовестного использования ИИ

Данная оценка построена на основе результатов экспертного опроса

Выборка и процедура проведения опроса

Всего в исследовании приняли участие (заполнили анкету опроса и согласились на интервью) 36 экспертов. Ниже представлено распределение экспертов, принявших участие в опросе (N=33), по категориям/сферам деятельности

Таблица 3

Выборка экспертов в рамках опроса

№	Сферы деятельности	Число экспертов в выборке
1	Представители ИТ-индустрии	7
2	Собственники, учредители (руководители) интернет-ресурсов	1
3	Представители онлайн-медиа, журналисты-обозреватели	2
4	Представители регулирующих/инвестиционных/ управляющих структур	6
5	Представители исследовательского/ академического сообществ	17
	Итого	**33**

Опрос реализован на платформе SurveyStudio в формате онлайн с 16 по 26 сентября 2024 года. Индивидуальные ссылки на заполнение анкеты рассылались по электронным адресам и/или номерам телефонов экспертов. Эксперты заполняли анкету самостоятельно.

Оценка остроты угроз, связанных с распространением дипфейков, и востребованности решений по их предотвращению

Для оценки остроты угроз, связанных с распространением дипфейков, экспертам был предложен список из 15 позиций, определенных на этапе кабинетного исследования, по каждой из которых предлагалось

дать оценку по 4-балльной шкале (от «0» до «3») со следующими значениями: «0» – угроза не представляет опасности, «1» – потенциальная опасность, «2» – высокий уровень опасности, «3» – крайне высокий уровень опасности.

Угрозы, предложенные для оценивания, разделены на 5 групп:

Группа 1: Дезинформация и манипуляция общественным мнением
- Использование дипфейков для дискредитации демократических институтов и процедур (например, выборов).
- Использование дипфейков для дискредитации судопроизводства (фальсификация улик, показаний).
- Производство фейковых новостей (fake news) с использованием дипфейков.
- Использование дипфейков в военных конфликтах (например, для дезориентации противника).

Группа 2: Киберпреступность
- Использование дипфейков как инструмента идентификации (KYC) для несанкционированного доступа к закрытым сервисам, включая банковские услуги.
- Использование дипфейков для фишинга (несанкционированный сбор информации, включая персональные данные).
- Использование дипфейков как инструмента финансового мошенничества (например, замена личности, просящей в долг).

Группа 3: Нарушение личных границ, приватности
- Дискредитация личностей посредством дипфейков в корпоративной сфере (создание синтетических сотрудников или копий существующих сотрудников).
- Дискредитация личностей в политической сфере, лидеров мнений, знаменитостей.
- Дипфейк-порнография (создание контента непристойного характера с использованием образа личности без его/ее согласия).

Группа 4: Угрозы государственной, корпоративной безопасности
- Использование дипфейков как инструмента шантажа личности и вымогательства.

- Распространение практик кибербуллинга с использованием дипфейков среди детей и подростков.
- Использование дипфейков как инструмента эскалации напряжения, разжигания и провоцирования конфликтов на различных уровнях, в т. ч. международном.

Группа 5: Негативные экономические последствия распространения дипфейков
- Использование дипфейков как инструмента для манипуляции рынками.
- Использование дипфейков для подрыва доверия к компаниям.

Полученные результаты отражены на рис. 4. Из него следует, что средние значения между 2 и 3 (то есть между высоким и крайне высоким уровнем опасности) эксперты присвоили 5 позициям, причем все они по своему значению ближе к высокому уровню (то есть менее 2,5). Остальные 10 позиций получили средние оценки между 1 (потенциальная опасность) и 2 (высокий уровень опасности), причем в 2 случаях опасность оценивается скорее как потенциальная, а в 8 – скорее как высокая.

В наименьшей степени эксперты видят угрозу в связи с возможностью эскалации напряжения, разжигания и провоцирования конфликтов.

Обратим внимание, что первые две позиции по уровню остроты занимают угрозы из категории «Дезинформация и манипуляция общественным мнением»: производство фейковых новостей и использование дипфейков в военных конфликтах. Далее следуют две позиции из категории «Киберпреступность»: фишинг и финансовое мошенничество. В наименьшей степени эксперты видят угрозу в связи с возможностью эскалации напряжения, разжигания и провоцирования конфликтов (категория «Угрозы государственной, корпоративной безопасности»), а также потенциальным подрывом доверия к компаниям (категория «Негативные экономические последствия распространения дипфейков»).

Рисунок 4

Экспертные оценки остроты угроз, связанных с распространением дипфейков (средний балл по шкале от 0 до 3)

Аналогичный список из 15 позиций был представлен во втором вопросе, в котором экспертам предлагалось выбрать решения по предотвращению угроз использования технологий дипфейков, которые будут наиболее востребованы и обеспечены наибольшим платежеспособным спросом на российском рынке в ближайшее время (1–2 года). Можно было выбрать до 5 позиций. Распределение ответов представлено на рис. 5.

В наименьшей степени эксперты оценивают востребованность и платежеспособный спрос на инструменты, направленные на предотвращение распространения практик кибербуллинга.

В данном случае очевидное лидерство принадлежит позиции «Предотвращение использования дипфейков как инструмента идентификации (KYC) для несанкционированного доступа к закрытым сервисам, включая банковские услуги» (82%; 27 выборов); на втором месте, также со значительным отрывом, позиция «Предотвращение использования дипфейков как инструментов финансового мошенничества» (64%; 21). Еще 3 позиции получили более трети экспертных голосов: «Предотвращение производства фейковых новостей (fake news) с использованием дипфейков» (45%; 15), «Предотвращение использования дипфейков для фишинга (несанкционированный сбор информации, включая персональные данные)» (39%; 13) и «Предотвращение использования дипфейков в военных конфликтах (например, для дезориентации противника)» (33%; 11).

В наименьшей степени эксперты оценивают востребованность и платежеспособный спрос на инструменты, направленные на предотвращение распространения практик кибербуллинга с использованием дипфейков среди детей и подростков (6%; 2), предотвращение использования дипфейков как инструмента для манипуляции рынками (9%; 3) и предотвращение использования дипфейков как инструмента эскалации напряжения, разжигания и провоцирования конфликтов на различных уровнях, в т. ч. международном (9%; 3).

Рисунок 5

Экспертные оценки востребованности решений по предотвращению угроз использования технологий дипфейков (%)

Оценка способов борьбы с дипфейками и профилактики их распространения

Существуют три принципиально различных направления борьбы с дипфейками. Во-первых, можно бороться с производством дипфейков посредством контроля за распространением и использованием соответствующих технологий. Во-вторых, можно решать вопрос на уровне каналов распространения дипфейков путем отслеживания, модерации и блокировки их размещения в социальных медиа. Наконец, в-третьих, можно развивать навыки и инструменты цифровой гигиены на уровне пользователей, повышая распознавание дипфейков и критическое отношение к медиаконтенту.

Экспертам было предложено проранжировать три перечисленных направления в зависимости от уровня их эффективности и перспективности. Чаще всего – 22 раза из 33 возможных – первое место было отдано цифровой гигиене. Ограничения распространения дипфейков на уровне каналов и контроль за распространением и использованием технологий рассматриваются в качестве менее эффективных, при этом перспективность данных направлений находится примерно на одном уровне (см. таблицу 4).

Таблица 4

Оценки эффективности трех направлений борьбы с дипфейками

	Ранг 1	Ранг 2	Ранг 3
Ограничение производства дипфейков (контроль за распространением и использованием технологий создания дипфейков)	5	**15**	13
Ограничение распространения дипфейков через каналы социальных медиа (отслеживание, модерация, блокировка)	6	**15**	12
Ограничение потребления дипфейков через развитие навыков и инструментов цифровой гигиены на уровне пользователей (распознавание и критическое отношение к медиаконтенту)	**22**	3	8

Следующий вопрос анкеты был посвящен способам борьбы с дипфейками и профилактики их негативного влияния. Таблица содержала 13 альтернатив, выявленных в ходе кабинетного исследования, которые условно разделены на 5 групп.

Группа 1: Разработка технологий для обнаружения дипфейков
- Создание и использование инструментов для выявления дипфейков для массового пользователя.
- Создание и использование профессиональных инструментов для выявления дипфейков (для журналистов, специалистов по безопасности, юристов).

Наиболее эффективными способами борьбы с угрозами, вызванными распространением дипфейков, оказываются меры цифровой гигиены на индивидуальном уровне.

Группа 2: Правовое регулирование
- Введение законодательной ответственности за использование дипфейков в деструктивных целях.
- Введение на законодательном уровне необходимости маркировки синтетического контента.
- Разработка и внедрение корпоративных правил и мер обеспечения безопасности клиентов (например, для банков, телекоммуникационных компаний).
- Разработка и внедрение корпоративных правил и мер по обеспечению внутренней безопасности сотрудников.

Группа 3: Образование и повышение осведомленности
- Развитие индивидуальных навыков критического восприятия информации.
- Следование рекомендациям по защите личной информации (двухфакторная идентификация, регулярная смена паролей, использование надежных паролей).
- Мероприятия по обучению цифровой гигиене уязвимых категорий пользователей.
- Мероприятия по пропаганде основных правил цифровой гигиены.

Группа 4: Этические стандарты и саморегулирование
- Разработка и внедрение этических кодексов, регламентов использования синтетического контента СМИ.
- Разработка и внедрение этических стандартов и практик саморегулирования компаниями, работающими с технологиями искусственного интеллекта, для предотвращения злоупотреблений дипфейками.

Группа 5: Сотрудничество между государственными и частными секторами
- Углубление сотрудничества между правительствами, технологическими компаниями и общественными организациями в области обмена информацией и разработки совместных решений по борьбе с дипфейками.

Для оценки эффективности перечисленных способов борьбы с дипфейками и профилактики их негативного влияния на сегодня использовалась шкала от «1» до «10» («1» – крайне низкий уровень эффективности, «10» – чрезвычайно высокий уровень эффективности).

Источник: freepik.com

Рисунок 6

Экспертные оценки эффективности способов борьбы с угрозами, вызванными распространением дипфейков

В качестве наиболее эффективных способов борьбы с дипфейками эксперты назвали развитие индивидуальных навыков критического восприятия информации (7,9) и создание и использование профессиональных инструментов для выявления дипфейков (7,7).

Наименее эффективные способы среди предложенных в анкете – разработка и внедрение этических кодексов, регламентов использования синтетического контента СМИ (4,5) и разработка и внедрение этических стандартов и практик саморегулирования компаниями, работающими с технологиями ИИ (4,7); обе позиции относятся к группе «Этические стандарты и саморегулирование». К ним примыкает единственная позиция из группы 5 «Сотрудничество между государственными и частными секторами», получившая среднюю оценку 5,5.

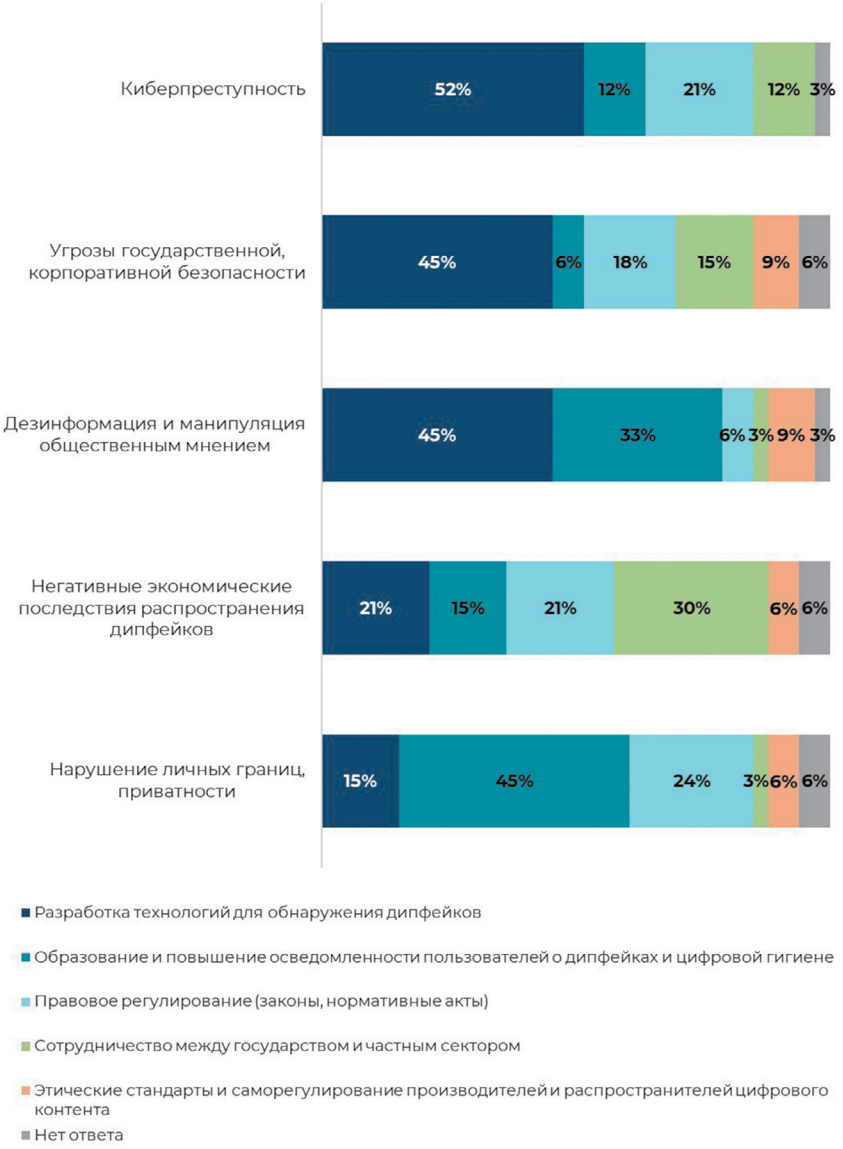
Наиболее эффективным способом борьбы с киберпреступностью и угрозами государственной и корпоративной безопасности эксперты считают разработку технологий для обнаружения дипфейков.

Таким образом, наиболее эффективными с точки зрения экспертов способами борьбы с угрозами, вызванными распространением дипфейков, оказываются меры цифровой гигиены на индивидуальном уровне (ограничения которых обсуждались в третьем параграфе Главы 2). Эксперты высоко оценивают значимость таких мер, однако массовые способы их практического применения ещё предстоит выработать.

Еще один вопрос был направлен на выявление соотношения между способами борьбы и профилактики для угроз разного типа. В данном случае в строках были представлены 5 групп угроз, а в столбцах – 5 групп решений. Предлагалось выбрать одно основное решение в каждой строке. Результаты отражены на рис. 7.

Рисунок 7

Экспертные оценки способов борьбы/профилактики для угроз разных типов, связанных с распространением дипфейков (%)



Как следует из данных на рис. 7, наиболее эффективным способом борьбы с киберпреступностью и угрозами государственной и корпоративной безопасности эксперты считают разработку технологий для обнаружения дипфейков. Данный способ также эффективен для борьбы с дезинформацией и манипуляцией общественным мнением – наряду с образованием и повышением осведомленности пользователей. Последняя мера рассматривается в качестве самой эффективной для профилактики нарушений личных границ, приватности. Наконец, с возможными негативными экономическими последствиями распространения дипфейков лучше всего бороться посредством сотрудничества между государственным и частным секторами.

Программные решения и организации, вовлеченные в борьбу с дипфейками в России

Экспертам были заданы открытые вопросы, в которых предлагалось назвать три наиболее эффективные программные разработки, отечественные или зарубежные, предназначенные для борьбы с угрозами дипфейков, а также конкретные институты и организации (фонды, ассоциации, ведомства, компании, платформы, сообщества), которые сегодня вносят наибольший вклад в борьбу с дипфейками в РФ.

При ответе на первый вопрос свои варианты программных решений указали 9 из 33 респондентов. Всего было названо 14 разработок, причем 8 из них фигурировали всего единожды, что может свидетельствовать о некотором номинальном разнообразии существующих технических решений в области борьбы с дипфейками (их эффективность относительно друг друга при этом не оценивается).

По 4 раза были упомянуты Microsoft Sentinel и FakeCatcher (Intel), 3 раза – Sensity AI, дважды – Deepware Scanner, WeVerify, Attestiv. Однократно были названы Detect Fakes (kellogg), Ghiro, TensorFlow, «Зефир», Microsoft Video Authenticator Tool, Project Origin, OzForensics, PyTorch. Стоит отметить, что большинство продуктов, названных экспертами, разрабатываются американскими (8) и (реже) европейскими компаниями (3). Среди отмеченных продуктов только два являются российскими разработками («Зефир», OzForensics). Примечательно, что даже в условиях санкций российские пользователи имеют доступ к зарубежным разработкам по противодействию дипфейкам. В таблице 5 представлена краткая информация о программных решениях, упомянутых экспертами.

Таблица 5

Наиболее эффективные программные разработки, предназначенные для борьбы с угрозами дипфейков, по мнению экспертов

Кол-во упоминаний	Название	Разработчик, страна	Аудитория	Описание способа/подхода к решению проблемы
4	Microsoft (Azure) Sentinel	Microsoft, США	B2C	Сбор данных облачного хранилища (о пользователе, устройстве, приложениях); выявление и аналитика угроз с помощью технологий ИИ, оперативное реагирование на инциденты с помощью встроенных средств координации и автоматизации стандартных задач
4	Fake-Catcher	Intel, США	B2B	Технология по различению в режиме реального времени «реального» человека на видео от подделки (дипфейка). С помощью ряда датчиков она отслеживает изменение цвета лицевых вен под кожей участника видео, создавая пространственно-временные карты. Затем алгоритмы машинного обучения по этим картам определяют, настоящий ли человек в кадре
3	Sensity AI	Sensity.ai, Нидер-ланды	B2B	Технология по выявлению дипфейков с помощью применения ИИ. Компания обещает обеспечить безопасное использование браузера для компаний и обучить сотрудников компаний распознавать дипфейки
2	Deepware Scanner	Zemana, Турция	B2C	Технология по обнаружению манипуляций с лицом, сгенерированных ИИ, в рамках видео (максимум 10-минутного), выложенного на YouTube, Facebook или Twitter.

Продолжение таблицы 5

Кол-во упоминаний	Название	Разработчик, страна	Аудитория	Описание способа/подхода к решению проблемы
2	WeVerify (the InVID-WeVerify verification plugin)	WeVerify Horizon Europe, Болгария	B2B	Платформа для совместной, децентрализованной проверки, отслеживания и разоблачения контента. Платформа с открытым исходным кодом, что позволяет привлекать сообщества и журналистов; обеспечивает интеграцию с внутренними системами управления контентом
1	Detect Fakes (kellogg)	Northwestern University (the Kellogg School of Management), США	B2C	Онлайн-тренажёр для проверки способности отличить настоящие изображения от изображений, сгенерированных ИИ
1	Ghiro	Ghiro, Италия	B2C	Автоматизированный инструмент, предназначенный для проведения криминалистического анализа большого количества изображений с помощью веб-приложения. Если нужно выполнить поиск по некоторым метаданным в группе изображений, определить местоположение группы изображений и просмотреть их на карте
1	Tensor-Flow	Google, США	B2B	Библиотека для машинного обучения с открытым исходным кодом для построения и тренировки нейронной сети с целью автоматического нахождения и классификации образов, достигая качества человеческого восприятия. Используется для обработки изображений, обработки естественного языка, голосовой и речевой обработки, анализа данных
1	«Зефир»	АНО «Диалог Регионы», Россия	B2B	Программа работает как информационная система мониторинга аудиовизуальных материалов на основе транскрибации в режиме реального времени. Определяет дипфейки благодаря алгоритмической оценке и анализу с помощью ИИ

Окончание таблицы 5

Кол-во упоминаний	Название	Разработчик, страна	Аудитория	Описание способа/подхода к решению проблемы
1	Attestiv	Attestiv, США	B2B	Автоматизированная проверка и анализ фотографий, видео и документов для компаний
1	Microsoft Video Authenti-cator Tool	Microsoft, США	B2C	Технология для выявления дипфейков – картинок или видео, которые создали с помощью компьютера, на которых изображение одного человека заменили лицом другого
1	Project Origin	Microsoft (Ignite) и другие, США	B2C	Сообщество, которое формирует список проверенных новостных провайдеров
1	Oz Foren-sics	Oz Forensics,	B2B	Платформа для проведения цифровых экспертиз, позволяющая распознать цифровые подделки фотографий и скан-копий документов (паспорт, ID, права, справки и другие документы) и обезопасить удалённую верификацию клиентов
1	PyTorch	Linux Foundation (до этого Facebook AI), США	B2B	Фреймворк для языка программирования Python, предназначенный для машинного обучения. Включает в себя набор инструментов для работы с моделями, используется в обработке естественного языка, компьютерном зрении и других похожих направлениях

Содержательный ответ на вопрос об организациях дали 17 из 33 экспертов, однако в некоторых случаях респонденты указали, что такие структуры им неизвестны.

У экспертов нет единого сложившегося представления о том, какие стейкхолдеры в настоящее время в России являются основными агентами в процессе предотвращения недобросовестного использования ИИ.

Всего было указано 18 организаций, в их число входят органы государственной власти, НКО, коммерческие компании и университеты. Три организации были названы по 4 раза: Сбер, Лаборатория Касперского и МВД. Трижды были названы «Яндекс» и Государственная Дума. По 2 раза в ответах экспертов фигурировали АНО «Диалог», АНО «Диалог Регионы», Минцифры и Альянс в сфере ИИ. Наконец, однократно упоминались: VisionLabs, Infowatch, Роскомнадзор, ФСБ, «Крибрум», СПб ФИЦ РАН, ИТМО, ДГТУ, Национальная федерация музыкальной индустрии.

В целом, такие результаты говорят о том, что у экспертов нет единого сложившегося представления о том, какие стейкхолдеры в настоящее время в России являются основными агентами в процессе предотвращения недобросовестного использования ИИ. Возможно, это связано с тем, что политика в части борьбы с дипфейками ещё только формируется, и число ситуативно вовлеченных в этот процесс стейкхолдеров оказывается достаточно велико.

Список потенциальных сервисов и продуктов, помогающих реализации практик цифровой гигиены

Согласно консолидированному мнению, выраженному участниками исследования, основная надежда по минимизации угроз, возникающих по мере распространения синтетического контента в целом и дипфейков в частности, связана с распространением правил цифровой гигиены и – шире – культуры поведения в цифровом пространстве.

Эксперты отмечают, что защитить персональную информацию в современном мире невозможно. Более того, по мере дальнейшего развития цифровых систем человек становится всё в большей степени информационно уязвимым.

«Я очень философски отношусь к вопросам утечки данных. Я считаю, что все данные рано или поздно утекут. Я считаю, что наступает эра абсолютной прозрачности всего и вся, и в этом смысле не сильно парюсь по поводу так называемых личных данных. Думаю, что все мои личные данные уже в куче самых разных баз, не очень понимаю, как это мне может угрожать, хотя, если хорошенько подумать, может быть масса разных способов. Но если речь идет о взятии кредитов на меня, там есть сейчас какие-то способы в банках, в Госуслугах, чтобы это законодательно тормозить. Всё остальное, то, что какие-то мои биологические данные, медицинские анализы – да ради бога, пускай они утекают, ничего страшного, так же, как паспортные данные, места жительства и т. д. Кому надо, тот и так всё установит, и камерами, если надо, всё проследит. Надо привыкать жить в эпоху абсолютной прозрачности».

О. А. Матвейчев, депутат Государственной Думы Федерального собрания Российской Федерации VIII созыва

Привыкание к новой информационной среде предполагает, что у пользователей должно выработаться восприятие ее как потенциально опасной, содержащей угрозы. В частности, в открытой цифровой среде перестают работать традиционные ценности журналистики факта. Если в классических СМИ публикуемый новостной материал по умолчанию является достоверным, то в цифровом информационном пространстве, предполагающем соседство профессионального контента с материалами,

генерируемыми пользователем, публикация является потенциально фейковой, так как большинство людей, публикующих контент, заведомо не проблематизируют критерий достоверности.

В целом, программные продукты, способные содействовать практикам цифровой гигиены, можно разделить на три группы.

1. Инструменты **«цифрового нотариуса»** – тип продуктов, направленных на поддержку проверки достоверности цифрового контента. Задача проверяющего в том, чтобы установить подлинность цифровых документов. Подобная функция может быть востребованной в широкой сфере профессиональных практик, охватывающих область традиционной нотариальной деятельности, однако ею не ограничивающихся. Например, сюда же можно отнести область проверки цифровых материалов журналистами или другими информационными работниками, занимающимися фактчекингом.

В качестве принципиального образца такого рода системы может рассматриваться программное обеспечение по определению антиплагиата, широко используемое в системе образования. Важно подчеркнуть, что программное обеспечение при таких проверках не является субъектом, несущим ответственность за принимаемое решение.

Ответственность несет человек – преподаватель и/или эксперт. Периодически встречаются случаи, когда высокий процент заимствований, фиксируемый программным обеспечением, не является основанием для определения плагиата; и наоборот, плагиат вполне может быть найден в ситуации отсутствия «претензий» со стороны системы.

2. Инструменты **«цифрового телохранителя»** – тип продуктов, предназначенных для защиты от атак с использованием дипфейков. Подобные решения должны максимально охватывать среды, через которые может быть передан фейковый контент – мессенджеры, сервисы видеоконференций, электронной почты, социальные сети. Задача в том, чтобы предупредить пользователя о возможном синтетическом характере контента, а в определенных случаях заблокировать его.

Принцип действия таких решений может быть сопоставлен с принципом работы антивирусных программ, защищающих цифровые устройства от вредоносного кода. Разница в том, что в случае дипфейков атака может быть направлена как на информационные системы, так и непосредственно на пользователя. Наряду с возможными программными разработками эксперты отмечают необходимость широкого информирования пользователей о возможных информационных угрозах, связанных с синтетическим контентом, а также массовой пропаганды приемов цифровой гигиены.

3. Одним из элементов такой информационной кампании может стать разработка **«цифрового советника»** – централизованного сервиса, способного выдать практические рекомендации в случае возникновения тех или иных угроз информационного характера.

Наиболее уязвимых категорий для «дипфейковой атаки», согласно консолидированному мнению участников исследования, две – это подростки и люди старшей возрастной группы (60+). И те и другие в наименьшей степени обладают цифровой культурой и опытом, необходимыми для противостояния дипфейковым угрозам. В то же время коммерческий потенциал создания таких продуктов оценивается как крайне слабый.

Еще одна уязвимая категория – корпоративные пользователи, с ними связан особый набор технологий мошенничества.

«Сейчас самые «модные» дипфейки, когда начальство звонит и просит перевести денег. У нас в институте до сих пор продолжаются именно видеозвонки, где сгенерировано изображение и голос человека, и есть люди, которые до сих пор на это попадаются, несмотря на то, что это доктора наук, профессора и проч., переводят деньги. Это мое окружение, мой возраст и старше, эта проблема есть».

К. Р. Нигматуллина, доктор политических наук, профессор кафедры цифровых медиакоммуникаций СПбГУ

«Если исторически смотреть, то уже давно появившиеся фишинговые атаки оказываются достаточно эффективными. Я помню, что, когда я работал в большой компании, у нас проводилась такая необъявленная проверка сотрудников, когда всем разослали некое письмо, которое было оформлено в корпоративном стиле, и нужно было куда-то нажать и что-то сделать. И смотрели процент, сколько людей на это повелось. Достаточно много таких людей оказалось, возможно потому, что это корпоративная почта, и если письмо с корпоративного аккаунта, то кажется, что всё хорошо.
Но на самом деле нужно быть более бдительным, потому что очень легко нажать куда-то, не подумав, даже людям, которые в IT вроде бы хорошо разбираются. Фишинговая угроза – это первое. Потом угроза взломов всевозможных, она всегда остается, потому что программное обеспечение несовершенно. Если у вас есть какой-то уязвимый ноут или телефон, то на него можно установить программное обеспечение, которое будет записывать экран, записывать звук, куда-то передавать. Такого рода угрозы всегда остаются. Поэтому люди, которые заклеивают камеру пластырем, возможно, делают правильно. Другое дело, что еще лучше, если скрывать нечего».

Д. В. Сошников, кандидат физико-математических наук, доцент факультета компьютерных наук НИУ ВШЭ, доцент МАИ

Источник: freepik.com

Наиболее перспективной с точки зрения финансирования программных продуктов для борьбы с дипфейками эксперты фактически единогласно считают банковскую индустрию. Представители последней способны выступить в качестве заказчиков инновационных решений.

Нужно быть более бдительным, потому что очень легко нажать куда-то, не подумав, даже людям, которые в IT вроде бы хорошо разбираются.

Еще одной категорией компаний, заинтересованных в подобных решениях, являются ИТ и телекоммуникационный секторы. Для них такие сервисы необходимы в качестве сопутствующих продуктов, обеспечивающих безопасность и надежность основных коммуникационных платформ. В первую очередь здесь выделяются крупные цифровые экосистемы, нуждающиеся в комплексных решениях такого рода.

Определенные категории акторов заинтересованы в подобных продуктах, однако возможности самостоятельно инициировать их создание у них относительно низкие. Сюда относятся образовательные организации, правовые структуры, многие медиакоммуникационные компании. Разработка сервисов для них может либо рассматриваться как коммерческая задача второго этапа, либо как социальная задача, финансирование которых должно осуществляться из некоммерческих фондов.

В таблице 6 каждый из рассмотренных типов ПО соотнесён с основными правилами цифровой гигиены, описанными во втором параграфе Главы 2.

Таблица 6

Типы ПО, которые помогут в реализации практик цифровой гигиены

Цифровой нотариус	Цифровой телохранитель	Цифровой советник
И1: Критическое отношение ко всей информации от незнакомых источников	**И2:** Безопасность паролей: использование уникальных паролей и их регулярная смена	**И1:** Критическое отношение ко всей информации от незнакомых источников
РО1: Разработка законодательства, регулирующего использование ИИ-технологий (в том числе, маркировка контента, созданного ИИ)	**И4:** Осторожность в сети: открывать только проверенные ссылки, скачивать приложения из надежных источников	**И2:** Безопасность паролей: использование уникальных паролей и их регулярная смена
РО4: Поддержка исследований и разработок в создании технологических средств защиты от киберугроз	**И5:** Цифровое благополучие: контроль времени, проведенного в сети, и просматриваемого контента	**И3:** Защита личных данных: не выкладывать конфиденциальную информацию даже в закрытые источники
	О1: Внедрите политики использования личных устройств	**И4:** Осторожность в сети: открывать только проверенные ссылки, скачивать приложения из надежных источников
	О3: Использование систем безопасности: программные или аппаратные решения, защищающие сеть организации от внешних угроз	**И5:** Цифровое благополучие: контроль времени, проведенного в сети, и просматриваемого контента

Окончание таблицы 6

Цифровой нотариус	Цифровой телохранитель	Цифровой советник
	О5: Работа с третьими сторонами: проверка систем безопасности партнеров и заключение договоров о конфиденциальности	**О2:** Регулярное обучение и информирование сотрудников о типах угроз и способах защиты
	РО1: Разработка законодательства, регулирующего использование ИИ-технологий (в том числе, маркировка контента, созданного ИИ)	**О4:** Защита данных: создание копий важных данных и их шифрование, использование надежного ПО
	РО4: Поддержка исследований и разработок в создании технологических средств защиты от киберугроз	**РО1:** Разработка законодательства, регулирующего использование ИИ-технологий (в том числе, маркировка контента, созданного ИИ)
		РО2: Стимулировании сотрудничества между ведомствами, частными и государственными компаниями для борьбы с фейковой информацией и киберугрозами
		РО3: Образовательные инициативы для повышения осведомленности населения

Заключение

По результатам проведенного исследования можно зафиксировать следующие тезисы. Дипфейк представляет собой вид синтетического контента, разрабатываемого с использованием глубоких нейронных сетей и применяемого для различных целей: от реализации образовательных инициатив до манипулирования общественным мнением и распространения ложной информации.

Развитие дипфейков неразрывно связано с прорывом в области технологий машинного обучения в 2010-х, приведшем к появлению больших генеративных моделей в начале 2020-х годов и общедоступных приложений по редактированию изображений, аудио и текстов.

Применение этих технологий сопровождается возникновением таких рисков, как снижение доверия целевых аудиторий к информации из Интернета в целом.

Случившееся одновременное с этим развитие рекомендательных сервисов социальных сетей резко усилило процесс популяризации дипфейков в информационном пространстве.
При этом фиксируется разнообразие сфер использования дипфейков: от медиаполя до политической арены.

Всё чаще дипфейки стали применяться в рамках создания пользовательского и развлекательного контента. Тем не менее многие эксперты подчеркивают неоднозначные последствия от массового внедрения подобных цифровых продуктов.

С одной стороны, открываются новые возможности для оптимизации процессов в различных отраслях. Например, дипфейки активно используются для производства фильмов, рекламы, а также персонализации и автоматизации услуг.

С другой – применение этих технологий сопровождается возникновением таких рисков, как снижение доверия целевых аудиторий к информации из Интернета в целом, подрыв кибербезопасности, осуществление политической манипуляции.

Риски от деструктивного использования дипфейков усиливаются на фоне относительной неготовности общества к такого рода фальсификациям (в том числе недостаточных навыков критического отношения к информации в цифровой среде) и технической невозможностью контроля идентификации дипфейков.
В рамках проекта удалось классифицировать потенциальные риски от использования дипфейков по объектам манипуляции, отраслевой специфике и воздействию на разные целевые аудитории.

При описании последней типологии следует отметить, что практики деструктивного использования дипфейков в первую очередь затрагивают пользователей сети с низким уровнем цифровой и финансовой грамотности. Относительно низкая стоимость создания дипфейков сделала их доступным инструментом для совершения адресных мошеннических действий, на уровне обычных граждан/пользователей посредством выброса синтезированных фотографий в соцмедиа, фальсификации голосов и видеоматериалов.

Практики деструктивного использования дипфейков в первую очередь затрагивают пользователей сети с низким уровнем цифровой и финансовой грамотности.

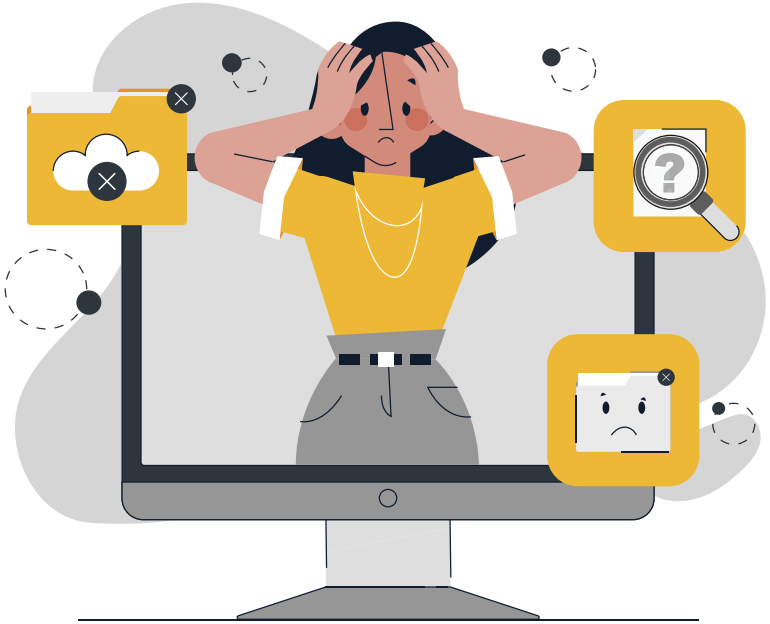
На федеральном уровне дипфейки становятся инструментом для манипуляций массовой аудиторией, формируя представления населения о политических деятелях, органах власти или брендах.

Таким образом, существенную роль в предотвращении рисков деструктивного использования дипфейков может сыграть массовое развитие навыков цифровой гигиены у всех социальных и профессиональных групп населения страны.

На сегодня основные сложности с освоением гражданами и компаниями набора цифровых навыков связаны с обобщенным характером самих профилактических рекомендаций, а также частыми трудностями в их исполнении.

Результаты опроса позволяют заключить, что из-за распространения дипфейков наиболее остро экспертное сообщество ощущает угрозы, связанные с манипуляцией общественным мнением и киберпреступлениями. В наименьшей степени угрозы связываются с негативными экономическими последствиями распространения дипфейков.

Наиболее востребованы на рынке решения, направленные на предотвращение использования дипфейков для обмана инструментов идентификации (KYC) для несанкционированного



Источник: freepik.com

доступа к закрытым сервисам, включая банковские услуги, и на предотвращение использования дипфейков в качестве инструментов финансового мошенничества.

Наиболее эффективным направлением борьбы с дипфейками эксперты считают развитие навыков и инструментов цифровой гигиены на уровне пользователей. В меньшей и примерно равной степени оценивается эффективность мер контроля за распространением и использованием дипфейк-технологий и отслеживания, модерации и блокировки размещений дипфейков в социальных медиа. При этом эксперты высоко оценивают важность создания и использования профессиональных инструментов для выявления дипфейков (для журналистов, специалистов по безопасности, юристов).

Эксперты высоко оценивают важность создания и использования профессиональных инструментов для выявления дипфейков

Самым действенным способом борьбы с киберпреступлениями и угрозами государственной/корпоративной безопасности, связанными с распространением дипфейков, эксперты считают разработку специальных инструментов для обнаружения данного типа синтетического контента.

Однако в настоящее время большинство экспертов затрудняются назвать эффективные сервисы и продукты, позволяющие бороться с дипфейками; распространена точка зрения, что надежность существующих разработок оставляет желать лучшего.

Исследование позволило выявить три класса информационных продуктов, направленных на предотвращение угроз, вызываемых распространением дипфейков. Это **«цифровой нотариус»** – продукты, предназначенные для выявления следов синтетического контента и поддержки верификации цифровых источников; **«цифровой телохранитель»** – инструменты для защиты от информационных атак

с использованием дипфейков; **«цифровой советник»** – сервисы, предоставляющие информационную поддержку, связанную с практиками цифровой гигиены.

Наиболее перспективные индустрии, способные поддержать разработку продуктов и сервисов для борьбы с угрозами дипфейков, – банковская и телекоммуникационная. Интерес, не поддержанный значительными финансовыми возможностями, есть у сферы образования, права и медиа. Наиболее уязвимые социальные группы – дети и подростки, старшие возрастные группы (60+), а также предприятия малого бизнеса.

Команда проекта

Сергей Алимбеков
заместитель директора
по технологическому
развитию

Кирилл Зендриков
директор департамента
экспертно-методического
сопровождения

Инна Скрытникова
директор департамента
по взаимодействию
с органами власти,
институтами развития
и экспертным сообществом

Мария Прокина
ведущий аналитик
департамента экспертно-
методического
сопровождения

Анастасия Агапова
аналитик департамента
экспертно-методического
сопровождения

Евгений Фелль
руководитель направления
по исследованию
и прогнозированию
информационных
технологий

Полина Шевелева
помощник аналитика

Ксения Тарабаева
помощник аналитика

Брошюра подготовлена командой ФРИИ на основе аналитического отчета
АНО «Социологическая мастерская Задорина» и собственных исследований Фонда.

В брошюре использованы результаты работ, выполненных сотрудниками ФРИИ
Е. Феллем, П. Шевелевой и К. Тарабаевой

В брошюре также использованы результаты работ,
выполненных командой АНО «Социологическая мастерская Задорина»
(Исследовательская группа ЦИРКОН) в составе:

Научный руководитель – И. В. Задорин
Руководитель проекта, к.ф.н. – С. Г. Давыдов
Заместитель руководителя – А. В. Сапонова
Приглашенный исследователь, к.э.н. – Н. Н. Матвеева
Исследователь – В. А. Чаленко
Исследователь – А. Л. Юргелас

Игорь Вениаминович Задорин
научный руководитель

Сергей Геннадьевич Давыдов
руководитель проекта, к.ф.н.

Анастасия Владимировна Сапонова
заместитель руководителя

Наталья Николаевна Матвеева
приглашенный исследователь, к.э.н.

Варвара Алексеевна Чаленко
исследователь

Анна Леонидовна Юргелас
исследователь

РЕДАКЦИЯ

Александр Василевский, редактор
Дмитрий Бабёнышев, дизайнер
Марина Гаева, корректор

спринт фрии

Минцифры России

sprint.iidf.ru

За более подробной информацией, пожалуйста,
обращайтесь по электронному адресу:

partners@iidf.ru
или по тел.
+ 7 495 258-88-77

Подготовлено в рамках реализации федерального проекта
«Цифровые технологии» национальной программы
«Цифровая экономика Российской Федерации»

ФРИИ, 2024