Давыдов Сергей Геннадьевич, кандидат философских наук, доцент департамента социологии, старший научный сотрудник Международной лаборатории исследований социальной интеграции, аналитик Международной лаборатории прикладного сетевого анализа, Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия; e-mail: sdavydov@hse.ru, ORCID 0000-0001-8455-9976

Матвеева Наталия Николаевна, кандидат экономических наук, доцент кафедры экономической теории и эконометрики, научный сотрудник Международной лаборатории прикладного сетевого анализа, Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия; e-mail: nmatvveva@hse.ru, ORCID 0000-0002-6378-7088

Сапонова Анастасия Владимировна, заместитель руководителя Исследовательской группы ЦИРКОН; старший преподаватель кафедры анализа социальных институтов, Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия; e-mail: saponova@zircon.ru, ORCID 0000-0002-9393-3509

ДИПФЕЙКИ КАК СОЦИАЛЬНАЯ УГРОЗА[1]

В статье дипфейки — аудио- и/или визуальный синтетический контент, создаваемый с помощью глубоких нейронных сетей, — рассматривается с позиций деструктивного использования. Активное развитие технологий дипфейков, а также рост числа мошенничеств с использованием этих технологий актуализируют потребность в систематической оценке возникающих социальных угроз. Авторами дается оценка возможных рисков для разных отраслей и сфер деятельности (политики, медиа, бизнеса и т. д.), а также для социального и психологического благополучия общества. На основании результатов опроса российских экспертов (N=33) определяются основные угрозы распространения дипфейков и уровень эффективности различных способов борьбы с ними. Также систематизируются практики цифровой гигиены как одного из наиболее эффективных методов профилактики указанных угроз для трех целевых групп: индивидов, организаций, регулирующих органов. В заключении представлена разработанная авторами типология программных продуктов, которые могут содействовать в реализации практик цифровой гигиены: цифровой «нотариус», цифровой «советник», цифровой «телохранитель».

***Ключевые слова**: дипфейк, искусственный интеллект, цифровая гигиена, цифровая безопасность.*

Как цитировать: *Давыдов С. Г., Матвеева Н. Н., Сапонова А. В.* Дипфейки как социальная угроза // Вестник Московского университета. Серия 10. Журналистика. 2025. № 5. С. 54–78. DOI: 10.55959/msu.vestnik.journ.5.2025.5478

Sergey G. Davydov, Candidate of Philosophical Sciences, Associate Professor, Department of Sociology, Senior Researcher, International Laboratory for Social Integration Research, Analyst, International Laboratory for Applied Network Research, HSE University, Moscow, Russia; e-mail: sdavydov@hse.ru, ORCID 0000-0001-8455-9976

Natalia N. Matveeva, Candidate of Economic Sciences, Associate Professor, Department of Economic Theory and Econometrics, Researcher, International Laboratory for Applied Network Research, HSE University, Moscow, Russia; e-mail: nmatvveva@hse.ru, ORCID 0000-0002-6378-7088

Anastasia V. Saponova, Deputy Director, ZIRCON Research Group; Senior Lecturer, Department of Social Institutions Analysis, HSE University, Moscow, Russia; e-mail: saponova@zircon.ru, ORCID 0000-0002-9393-3509

DIPFAKES AS A SOCIAL THREAT

The article examines the phenomenon of deepfakes – audio and/or visual synthetic content created using deep neural networks – from the perspective of its destructive use. The rapid development of deepfake technologies and the increasing number of scams using them make it necessary to systematically assess emerging social threats. The authors attempt to provide an assessment of the risks to various industries and fields, such as politics, media, business, and social and psychological well-being. Based on a survey of 33 Russian experts, the main threats to the spread of deepfakes and the effectiveness of methods for combating them were identified. Digital hygiene practices, which are recognized as the most effective method of preventing these threats, have also been systematized in order to prevent threats to three target audiences: individuals, organizations and governments. Finally, the authors present a typology of software products for implementing digital hygiene, including a digital notary, digital adviser, and digital bodyguard.

***Key words**: deepfake, artificial intelligence, digital hygiene, digital security.*

To cite this article: Davydov S. G., Matveeva N. N., Saponova A. V. (2025) Dipfeyki kak sotsial'naya ugroza [Dipfakes as a social threat]. *Vestnik Moskovskogo Universiteta. Seriya 10. Zhurnalistika* 5: 54–78. DOI: 10.55959/msu.vestnik.journ.5.2025.5478

DOI: 10.55959/msu.vestnik.journ.5.2025.5478

Что такое дипфейк?

В последние годы слово «дипфейк» получило широкое распространение. Согласно распространенной версии, в английском языке оно появилось в 2017 г. благодаря пользователю онлайн-форума *Reddit*, который выкладывал в папку с таким названием поддельные порноролики «с участием знаменитостей»[2]. Слово

возникло путем сложения двух терминов. Первый – *deep learning* (глубокое обучение). Это разновидность машинного обучения, обучение искусственных нейронных сетей на больших объемах данных. Второй – *fake* (подделка или фальшивка). Следует отметить, что русская калька «фейк» имеет более узкое значение. Это не всякая фальшивка, но информационная (*fake news*).

Дипфейки – частный случай так называемого синтетического контента, то есть медийного контента, созданного с использованием глубоких нейронных сетей (*DNN*) (Mirsky, Lee, 2021). Конвенционально признано, что синтетический контент имеет широкий потенциал для практического использования, в частности в развлекательных и образовательных целях. Известные примеры добросовестного использования технологии дипфейков – отрисовка образов актера Брюса Уиллиса для рекламы Мегафона или персонажа советской кинокомедии «Иван Васильевич меняет профессию» Жоржа Милославского (актер Леонид Куравлев) для рекламы Сбера[3]. В то же время активно обсуждаются широкие возможности использования дипфейков в деструктивных целях, о которых подробнее будет сказано ниже.

В российской научной и научно-практической литературе дипфейки рассматриваются прежде всего в правовом контексте. В. Г. Иванов и Я. Р. Игнатовский определяют их как «методику компьютерного синтеза изображения, основанную на искусственном интеллекте, которая используется для соединения и наложения существующих изображений и видео на исходные изображения или видеоролики» (2020: 380–381). По мнению А. С. Киселева, «дипфейками называют реалистичную замену лиц и голоса посредством использования генеративно-состязательных нейросетей» (2021: 55–56). Формулировка, которую предлагают Н. Р. Красовская и А. А. Гуляев, – «совокупность технологических трансформаций изображения и видео, созданных с использованием искусственного интеллекта» (2020: 93).

Определения дипфейков в научных публикациях, примеры которых приведены выше, близки и друг другу не противоречат. Их анализ позволяет заключить, что под дипфейком понимается как метод обработки аудио и/или графики посредством определенных технологий искусственного интеллекта, так и медийный контент, сгенерированный при помощи этого метода. И в том и в другом случае речь идет о нейтральном явлении, вне зависимости

от контекста использования. Слово «фейк», однако, и в русском, и в английском языках имеет явно выраженную негативную коннотацию.

Мы предлагаем использовать термин «дипфейк» для обозначения цифрового медийного контента, сгенерированного посредством обработки аудио и/или графики при помощи глубоких нейронных сетей (*DNN*), использование которого имеет признаки недобросовестного.

За сравнительно недолгую историю существования дипфейки успели приобрести популярность в качестве объекта анализа. Результаты ряда исследований, посвященных способности идентификации дипфейков разных типов (напр.: Bray, Johnson, Kleinberg, 2023; Köbis, Doležalová, Soraperra, 2021; Mai, Bray, Davies, Griffin, 2023), свидетельствуют о высоком уровне ошибок, допускаемых испытуемыми. Учитывая фактор совершенствования технологий, М. Хамелерс и соавторы (Hamelers, van der Meer, Dobber, 2023) делают вывод, что основным способом выявить дипфейк для аудитории, не обладающей специальными инструментами, станет сомнение в контекстуальной достоверности. Внимание исследователей также привлекают перспективы распространения технологий дипфейков и отношение к ним различных референтных групп. По результатам весьма содержательного исследования в 7 европейских странах, Н. Хайнек и соавторы (Hynek, Gavurova, Kubak, 2025) заключают, что в общественном мнении дипфейки устойчиво ассоциируются в большей степени с угрозами, чем с возможностями. К числу угроз чаще всего относят дезинформацию как оружие, нарушение границ приватности, «гонку вооружений» в области обнаружения дипфейков, а также малоизученное влияние в медицине, образовании и креативных индустриях.

Вместе с тем в экспертной среде распространено мнение о переоцененности дипфейков как носителей социальных угроз. Разделяя точку зрения о необходимости дальнейшего активного изучения соответствующего круга проблем, А. Бирер и Н. Джаст (Birer, Just, 2024) призывают опираться на исследовательские данные и факты в вопросах регулирования. Р. Кавер предлагает рассматривать дипфейки с позиций социокультурных исследований, акцентируя внимание на трансформации практик создания и интерпретации текстов, игровой и творческой деятельности (Cover, 2022).

Обзор зарубежных научных публикаций, направленных на осмысление дипфейков как социокультурного феномена, свидетельствует об актуальности проведения его комплексного анализа на российском материале. Данной статьей мы надеемся внести вклад в решение этой задачи.

Постановка проблемы и методика исследования

В российской научной литературе проблематика использования технологий дипфейков рассматривается преимущественно с теоретических позиций. К ней в большей степени обращаются юристы (в связи с лакунами в российском законодательстве, которые существуют вокруг использования технологий дипфейков), нежели исследователи социальной коммуникации. Число случаев мошенничества с использованием дипфейков продолжает расти[4], однако систематическая оценка угроз от распространения использования технологий дипфейков в российском контексте до настоящего времени не была представлена. Технологии дипфейков могут использовать с целью не только финансового мошенничества, но и политической манипуляции, дезинформации, несанкционированного доступа к банковским услугами и т. д. В этой связи представляется важным обратиться к экспертам из разных отраслей для получения представления о текущем уровне угрозы, которую несут дипфейки, а также возможным направлениям противодействия им. Таким образом, мы сформулировали следующие исследовательские вопросы:

– Как экспертное сообщество оценивает потенциальные риски и угрозы от распространения технологий дипфейков?

– Как можно оценить эффективность мер противодействия недобросовестному использованию дипфейков?

– Каким образом программные продукты могут содействовать практикам цифровой гигиены?

Далее мы представим результаты опроса российских экспертов об угрозах дипфейков и мерах борьбы с ними, систематизируем информацию об основных правилах цифровой гигиены как потенциально одного из самых эффективных способов борьбы с угрозами дипфейков, представим разработанную типологию программных продуктов, которые могут содействовать в реализации практик цифровой гигиены.

Исследование, результаты которого приводятся в настоящей статье, было реализовано Исследовательской группой ЦИРКОН (АНО «Социологическая мастерская Задорина») при поддержке Фонда развития интернет-инициатив (ФРИИ). Оно состояло из двух частей – экспертного опроса и экспертных интервью. В рамках экспертного опроса ставилась задача получить оценку текущих рисков, которые несут технологии дипфейков, и эффективности способов борьбы с обозначенными угрозами, а также востребованности отдельных технологических решений для предотвращения угроз использования технологий дипфейков. В ходе экспертных интервью фокусированно обсуждались вопросы, связанные с цифровой гигиеной, как ключевым способом предотвращения угроз, связанных с распространением дипфейков. В частности, обсуждались перспективные направления для развития технологий и сервисов, которые бы помогали пользователям соблюдать цифровую гигиену.

Экспертный опрос представлял собой самостоятельное заполнение количественной онлайн-анкеты на платформе *SurveyStudio*. На электронную почту и/или телефоны экспертов были разосланы индивидуальные приглашения со ссылками на анкету. В рамках второй части исследования были проведены экспертные интервью (формат — онлайн). Сбор данных происходил в период с 16 по 26 сентября 2024 г. Всего в исследовании приняли участие 36 экспертов. Было заполнено 33 анкеты и проведено 6 интервью; 3 эксперта участвовали как в опросе, так и в интервью. К участию в исследовании приглашались эксперты следующих пяти категорий (см. табл. 1).

Таблица 1

Выборка экспертного опроса

№	Сферы деятельности	Число экспертов в выборке
1	Представители ИТ-индустрии	7
2	Собственники, учредители (руководители) интернет-ресурсов	1
3	Представители онлай-медиа, журналисты-обозреватели	2

Окончание таблицы 1

4	Представители регулирующих/инвестиционных/ управляющих структур	6
5	Представители исследовательского/ академического сообществ	17
Итого		**33**

Экспертный опрос проектировался следующим образом: на основании информационного поиска был сформирован список российских экспертов, чье поле деятельности связано с вопросами информационной безопасности, использования технологий ИИ, создания синтетического контента и дипфейков, в частности. Для того чтобы собрать экспертные оценки от разных профессиональных групп, были обозначены пять категорий экспертов: ИТ-специалисты, менеджеры информационных ресурсов, журналисты, регуляторы, исследователи. Также принимались рекомендации по релевантным экспертам от самих экспертов, которые участвовали в опросе. Всего приглашения пройти опрос были направлены 99 экспертам, из них анкету заполнили 33. Стоит отметить, что ограничением настоящего исследования является опора в большей степени на публичных экспертов по обозначенной проблематике, т. е. мы получаем срез экспертных оценок от тех специалистов, которые наиболее видимы и потенциально достижимы. Ограничением является также тот факт, что наибольший отклик (ожидаемо) пришелся на категорию исследователей и представителей академического сообщества, которые составили почти половину итоговой выборки (17).

Количественная анкета состояла из вопросов, направленных на оценку угроз, связанных с развитием технологий дипфейков, и способов их предотвращения. Также был включен блок открытых вопросов с пожеланием указать наиболее эффективные инструменты борьбы с дипфейками, организации, занимающиеся такой борьбой, наиболее значимые тексты по обсуждаемому вопросу и их авторов.

Экспертная оценка угроз от использования дипфейков

Для оценки остроты угроз, связанных с распространением дипфейков, экспертам был предложен список из 15 позиций[5], 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. Эксперты оценивали остроту угроз на сегодняшний момент и применительно к российскому контексту.

Полученные результаты отражены на рисунке 1: 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 (менее 2,5). Остальные 10 позиций получили средние оценки между 1 (потенциальная опасность) и 2 (высокий уровень опасности), причем в двух случаях опасность оценивается скорее как потенциальная, а в 8 — скорее как высокая.

Обратим внимание, что первые две позиции по уровню остроты занимают угрозы производства фейковых новостей и использование дипфейков в военных конфликтах (т. е. связанные с дезинформацией и манипуляцией общественным мнением). Далее следуют фишинг и финансовое мошенничество, представляющие собой разные виды киберпреступности. В наименьшей степени эксперты видят угрозу в связи с возможностью эскалации напряжения, разжигания и провоцирования конфликтов, а также потенциальным подрывом доверия к компаниям.

Таким образом, результаты опроса показывают, что в связи с распространением 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 на уровне бизнеса, а также с провоцированием конфликтов на различных уровнях. Такое распределение во многом связано с уже существующими прецедентами использования дипфейков. Поэтому интерпретировать оценки угроз можно только как характеристику текущего контекста, без попытки экстраполяции даже на краткосрочную перспективу.



Рисунок 1. Экспертные оценки остроты угроз, связанных с распространением дипфейков (средний балл по шкале от 0 до 3)[6]

Аналогичный список из 15 позиций был представлен в следующем вопросе (см. рис. 2), 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.

В данном случае очевидное лидерство принадлежит позиции «Предотвращение использования дипфейков как инструмента идентификации *(KYC — Know Your Customer)* 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 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).

Таким образом, 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. 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 рассматриваться либо как коммерческая задача второго этапа, либо как социальная задача, финансирование которых должно осуществляться из некоммерческих фондов.

Наиболее уязвимых категорий для «дипфейковой атаки», согласно консолидированному мнению экспертов, 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, как уже упоминалось выше, оценивается как крайне слабый.

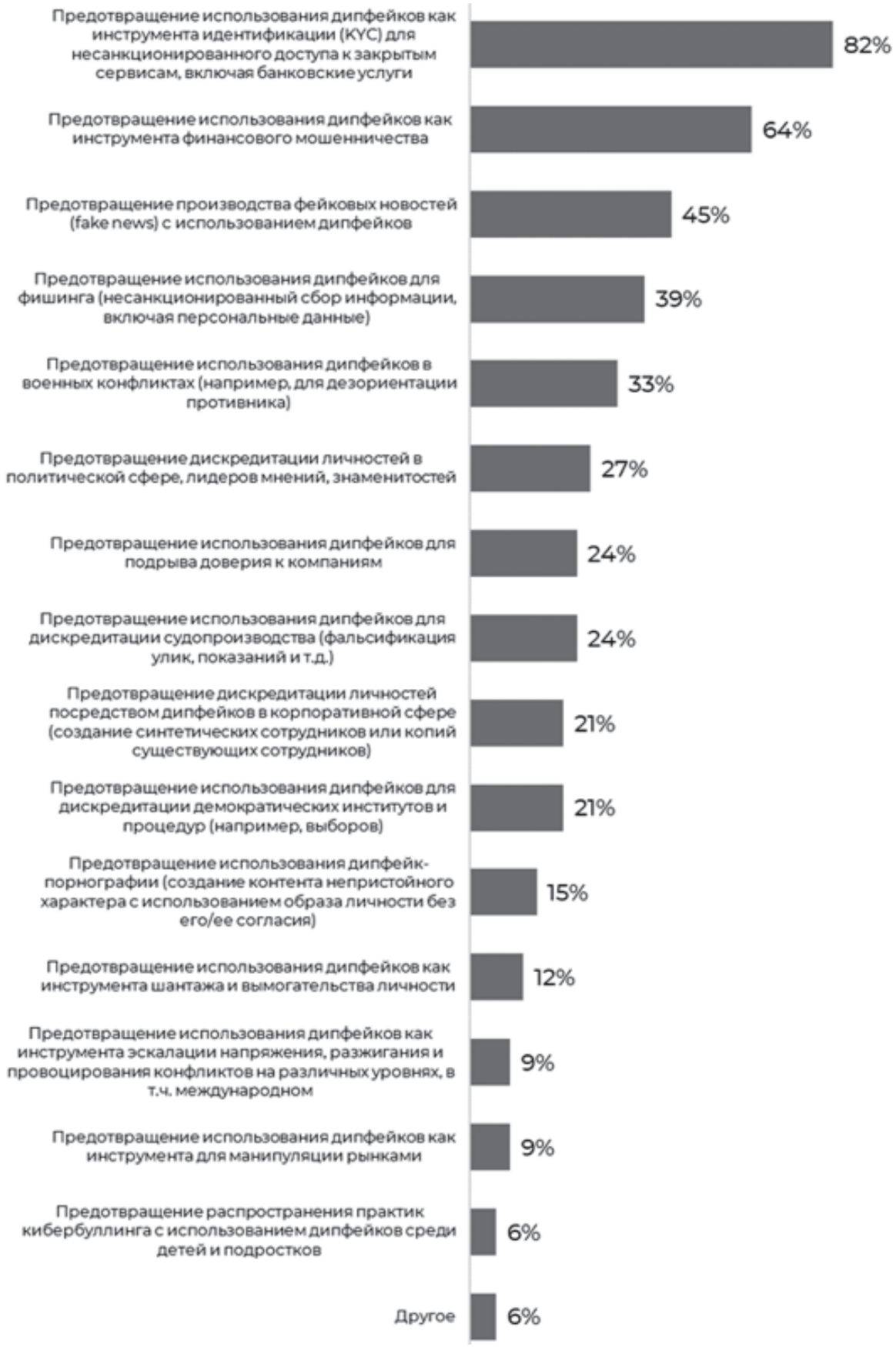


Рисунок 2. Экспертные оценки востребованности решений по предотвращению угроз использования технологий дипфейков (%)[7]

Следующий вопрос анкеты был посвящен способам борьбы с дипфейками и профилактике их негативного влияния[8]. Для оценки их эффективности на сегодняшний день использовалась шкала от «1» до «10» («1» — крайне низкий уровень эффективности, «10» — чрезвычайно высокий уровень эффективности).

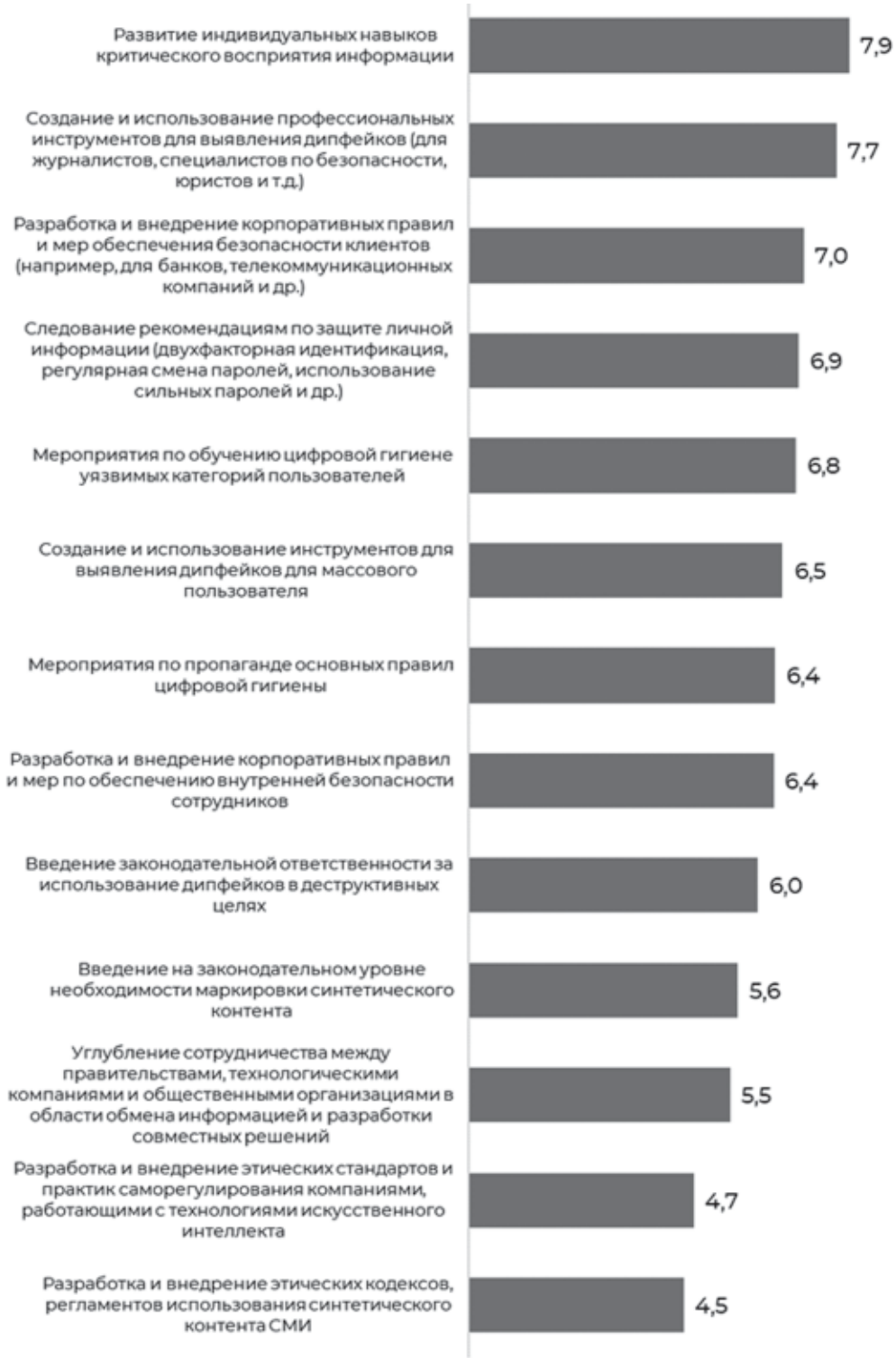


Рисунок 3. Экспертные оценки эффективности способов борьбы с угрозами, вызванными распространением дипфейков (средний балл по шкале от 1 до 10)[9]

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). Невысоко также была оценена эффективность сотрудничества между государственными и частными секторами (5,5), а также законодательное регулирование в части маркировки синтетического контента (5,6).

Таким образом, наиболее эффективными, с точки зрения экспертов, способами борьбы с угрозами, вызванными распространением дипфейков, оказываются меры цифровой гигиены на индивидуальном уровне, ограничения которых обсуждаются ниже. Эксперты высоко оценивают значимость таких мер, однако массовые способы их практического применения еще предстоит выработать. В этом могут помочь конкретные технологические решения, которые мы обсудим далее.

Цифровая гигиена как средство профилактики угроз, вызванных распространением дипфейков

С развитием информационных технологий способы мошенничества и обмана в цифровом пространстве становятся все более изощренными: создаются поддельные сайты, фальшивые новости, организации и люди. Поскольку технология дипфейка создается на передовой технологических изобретений, существующие законодательные и технические меры профилактики и борьбы с недобросовестным контентом зачастую устаревают и становятся малоэффективными. По этой причине специалисты в области кибербезопасности отмечают цифровую гигиену как основной способ предупреждения и борьбы с похищением данных и мошенничеством.

Цифровая гигиена в широком смысле — это набор навыков, помогающих избегать рисков, связанных с информационными технологиями (Суходаева, 2022). 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 достаточно давно, вместе с развитием цифровых ресурсов (телефоны, смартфоны, интернет). По аналогии с общепринятой гигиеной цифровая гигиена призвана сохранить здоровье и обеспечить благополучие человека в цифровом мире.

Меры цифровой гигиены направлены на предотвращение и профилактику угроз и рисков, возникающих в процессе использования электронных ресурсов. К основным угрозам, с которыми сталкиваются пользователи в интернете, можно отнести следующие: кража личных данных и информации третьими лицами; мошенничество; потеря психологического благополучия и ментального здоровья.

Базовые правила цифровой гигиены направлены на предотвращение утечки данных, защиту от мошенничества и сохранение психологического благополучия. К ним относятся как поведенческие навыки, направленные на защиту личной информации и контроль времени, проведенного в интернет-пространстве, так и технологические, предполагающие создание надежных паролей и использование двухфакторной аутентификации (Ашманов, Касперская, 2021) Данные меры являются универсальными, однако степень уязвимости пользователей отличается в зависимости от возраста, образования, видов и частоты использования цифровых ресурсов. По этой причине специалисты в области кибербезопасности рекомендуют индивидуально оценивать возможные риски и, по возможности, индивидуально подобрать программу и инструменты обучения цифровой гигиене. Среди наиболее уязвимых категорий отмечаются дети и лица пожилого возраста (Суходаева, 2022). С появлением дипфейков вопросы цифровой гигиены вышли на новый уровень.

Меры цифровой гигиены для защиты от дипфейков различаются в подходах и акцентах. Некоторые источники делают упор на использование передовых технологий для выявления дипфейков[10], другие[11] 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, включающие 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. В настоящее время ряд компаний активно занимаются разработкой ПО для обнаружения дипфейков. Например, проект *DARPA* под названием *SemaFor* разрабатывает инструменты для обнаружения манипуляций с медиаконтентом[12]. Другой пример: специалисты Института искусственного интеллекта *AIRI* и Московского технического университета связи и информатики (МТУСИ) разработали ИИ-модель для определения синтетических голосов, которые используют телефонные мошенники (Borodin, Kudryavtsev, Korzh, Efimenko et al., 2024).

Поведенческие рекомендации направлены на помощь пользователю в распознавании ложной информации и предотвращение обмана. К таким мерам относятся: защита личной информации и осторожность при предоставлении личной информации онлайн (не делиться критической информацией, избегать перехода по подозрительным ссылкам), загрузка приложений и файлов только из официальных источников, мониторинг и отслеживание упоминаний о себе. В данном подходе особое место уделяется образовательному аспекту. Ряд специалистов подчеркивает важность образования и повышения осведомленности о технологиях дипфейков не только на индивидуальном, но и на корпоративном уровне. Некоторые источники предлагают конкретные обучающие программы[13].

Отдельные источники затрагивают правовые и этические аспекты использования технологий дипфейков, предлагая рассмотреть вопросы регулирования. Среди регулирующих мер со стороны государства в первую очередь отмечается введение обязательств на указание способа производства контента, если он произведен с использованием нейронных сетей. На корпоративном уровне отмечается необходимость введения систем и правил, препятствующих мошенническим действиям в отношении потребителя. К примеру, ЦБ России регулярно разрабатывает рекомендации для банков с целью защиты клиентов от кражи их средств. Ряд компаний самостоятельно разрабатывают рекомендации для защиты от фейков и дипфейков. Например, аудиторская компания *KPMG* для защиты от дипфейков рекомендует компаниям уделить особое внимание внедрению инструментов кибербезопасно-

сти и введению системы изначального нулевого доверия к любой информации. В основе такой системы лежит развитие навыков цифровой гигиены как превентивного источника защиты от обмана и мошенничества.

Можно выделить следующие базовые рекомендации для индивидуальных пользователей, компаний и регулирующих органов (см. табл. 2). Пункты в таблице не ранжируются по важности.

Таблица 2

Основные правила цифровой гигиены

Для индивида	Для организации	Для регулирующих органов
Критическое отношение ко всей информации от незнакомых источников	Внедрите политики использования личных устройств	Разработка законодательства, регулирующего использование ИИ-технологий (в том числе, маркировка контента созданного ИИ)
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
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
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
Цифровое благополучие: контроль времени, проведенного в Сети, и просматриваемого контента	Работа с третьими сторонами: проверка систем безопасности партнеров и заключение договоров о конфиденциальности	

На индивидуальном уровне особое внимание уделяется навыкам недоверия незнакомым источникам и защите своих личных данных. 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 этого, 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. Причины таких решений связаны с опасениями, что зарубежные сервисы могут быть использованы для шпионажа или сбора конфиденциальной информации. Кроме этого, законодательство некоторых стран может требовать хранения данных на территории их страны.

На государственном уровне требуется введение регулирования размещения контента, созданного с помощью новых технологий. Кроме этого, требуют проработки вопросы утечки персональных данных, которые становятся материалом для мошенников. По словам Натальи Касперской, данные утекают в основном из-за сотрудников, работающих с персональными данными в банках, телекоммуникационных компаниях и государственных учреждениях. Эксперт считает, что решить эту проблему можно только изменив подход к ответственности за неправомерное использование персональных данных. Еще одним вопросом, нуждающимся в законодательной проработке, являются юридические особенности определения ответственности при атаках с использованием поддельных личностей (защита «оригинала», личность которого была украдена). Помимо законодательного регулирования, от государства требуется создание условий для обучения и проведения исследований.

Цифровая гигиена является важной профилактикой для защиты от киберугроз на разных уровнях, однако эффективность представленных выше рекомендаций имеет ряд ограничений. Во-первых, зачастую рекомендации носят общий характер (не открывать подозрительные ссылки, не доверять недостоверным источникам и т. д.). Данные рекомендации не учитывают конкретную ситуацию или

организацию, поэтому пользователь зачастую может не ассоциировать свою ситуацию с критической. Кроме того, степень доверия либо подозрения зависит от конкретного индивида, его психологических особенностей и опыта. Во-вторых, отдельные рекомендации могут противоречить друг другу или быть несовместимыми с определенными рабочими процессами. Например, некоторые источники рекомендуют использовать сложные пароли для каждого аккаунта и хранить их в «надежной» базе паролей. Другие советуют запоминать все пароли и не хранить их в едином файле. В-третьих, некоторые рекомендации являются сложными для понимания пользователей или реализации компанией ввиду высокой стоимости. Например, настройка двухфакторной аутентификации, шифрование данных и установка комплексных систем безопасности.

Тем не менее рекомендации важны для обеспечения базовой защиты и повышения общей осведомленности. Эффективность рекомендаций может быть выше, если излагать их более понятным и конкретным языком и проводить регулярное обучение пользователей навыкам цифровой гигиены. Кроме этого, автоматическое включение части систем безопасности может облегчить работу пользователей. Помимо этого, навыки цифровой гигиены и рекомендации требуют постоянного обновления с учетом новых угроз.

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

В рамках опроса экспертам было предложено назвать 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 России.

При ответе на первый вопрос свои варианты программных решений предложили 9 из 33 экспертов. 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. При этом отметим, что их эффективность относительно друг друга 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, 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 (однако в некоторых случаях этот доступ частичный).

Содержательный ответ на вопрос об организациях дали 17 из 33 экспертов, однако в некоторых случаях респонденты указали, что такие структуры им неизвестны. 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», «Санкт-Петербургский федеральный исследовательский центр 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 еще только формируется и число ситуативно вовлеченных в этот процесс стейхолдеров оказывается достаточно велико.

Заключение

Технологии генерации синтетического контента, включая инструменты создания дипфейков, меняют информационное пространство, становясь, если пользоваться терминологией М. Маклюэна (2007), очередным «посланием» (*message*) или вызовом, на который человечеству необходимо выработать ответ. В данном случае вызов заключается в выработке компетенций и создании инструментов, позволяющих отличить «документальный» контент от «художественного». Опыт развития коммуникационных техно-

логий и практик показывает, что на предыдущих исторических этапах эта задача неизменно решалась – от различения волшебных сказок и исторических нарративов до верификации информации в современной новостной индустрии.

Результаты проведенного исследования свидетельствуют о критической важности цифровой гигиены для предотвращения недобросовестного использования дипфейков. 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 заключается в том, чтобы установить подлинность цифровых документов. Подобная функция может быть востребована в широкой сфере профессиональных практик, охватывающих область традиционной нотариальной деятельности, однако ею не ограничиваться. Например, сюда 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 плагиата, 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 и др. Задача заключается в том, чтобы в определенных случаях заблокировать вредоносный контент.

Принцип действия таких решений может быть сопоставлен с принципом работы антивирусных программ, защищающих цифровые устройства от вредоносного кода. Разница заключается в том, что в случае дипфейков атака может быть направлена как на информационные системы, так и непосредственно на пользователя.

3) Инструмент «цифрового советника» – тип централизованного сервиса, способного выдать практические рекомендации в случае возникновения тех или иных угроз информационного характера.

В отличие от типа инструментов «цифровой телохранитель», чья функция скорее ограничительная, задача инструментов типа «цифровой советник» рекомендательная: они могут предупредить пользователя о возможном синтетическом характере контента, однако окончательное решение о действии в отношении обозначенной угрозы остается за пользователем.

Наряду с возможными программными разработками, необходимы широкое информирование пользователей о возможных информационных угрозах, связанных с синтетическим контентом, а также массовая пропаганда приемов цифровой гигиены.

Примечания

[1] Статья подготовлена в рамках Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики».

[2] Roettgers J. (2018) Porn Producers Offer to Help Hollywood Take Down Deepfake Videos. *Variety*. Available at: https://variety.com/2018/digital/news/deepfakes-porn-adult-industry-1202705749/ (accessed: 15.06.2025).

[3] *Воронин М. С.* Брюс Уиллис и Жорж Милославский: как создаются и используются дипфейки // Проектно-учебная лаборатория экономической журналистики НИУ ВШЭ. 11.04.2022. Режим доступа: https://economics.hse.ru/ecjourn/news/588939161.html (дата обращения: 15.06.2025).

[4] Эксперты сообщили о резком росте числа мошенничеств с использованием дипфейков // *Forbes*. Режим доступа: https://www.forbes.ru/tekhnologii/539953-eksperty-soobsili-o-rezkom-roste-cisla-mosennicestv-s-ispol-zovaniem-dipfejkov (дата обращения: 14.09.2025).

[5] Сформированная на этапе кабинетного исследования типология угроз представлена в Приложении 1.

[6] На диаграмме представлены средние арифметические значения по экспертным оценкам.

[7] Проценты рассчитаны от числа экспертов, принявших участие в опросе.

[8] Сформированная на этапе кабинетного исследования типология способов борьбы с распространением дипфейков представлена в Приложении 2.

[9] На диаграмме представлены средние арифметические значения по экспертным оценкам.

[10] Good Digital Hygiene Practices You Need to Know. *Vodafone*. Available at: https://www.vodafone.com.au/blog/security/good-digital-hygiene-practices-you-need-to-know (accessed: 15.06.2025); Semantic Forensics. *DARPA*. Available at: https://www.darpa.mil/program/semantic-forensics (дата обращения: 15.06.2025).

[11] Dartmouth Guide to Digital Hygiene. *Dartmouth College*. Available at: https://services.dartmouth.edu/TDClient/1806/Portal/KB/ArticleDet?ID=155669 (accessed: 15.06.2025); Digital Hygiene. *Group-IB*. Available at: https://group-ib.medium.com/digital-hygiene-338f4c60f24c (accessed: 15.06.2025).

[12] Dartmouth Guide to Digital Hygiene. *Dartmouth College*. Available at: https://services.dartmouth.edu/TDClient/1806/Portal/KB/ArticleDet?ID=155669 (accessed: 15.06.2025).

[13] Uncovering Deepfakes: Classroom Guide + Discussion Questions. *AI for Education*. Available at: https://www.aiforeducation.io/ai-resources/uncovering-deepfakes (accessed: 15.06.2025).

Библиография

Ашманов И. С., Касперская Н. И. Цифровая гигиена. СПб.: ИД «Питер», 2021.

Иванов В. Г., Игнатовский Я. Р. Deepfakes: перспективы применения в политике и угрозы для личности и национальной безопасности // Вестн. РУДН. Сер.: Государственное и муниципальное управление. 2020. Т. 7. № 4. С. 379–386.

Киселев А. С. О необходимости правового регулирования искусственного интеллекта: дипфейк как угроза национальной безопасности // Вестн. Моск. гос. обл. ун-та. Сер.: Юриспруденция. 2021. № 3. С. 54–64.

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.

Маклюэн М. Понимание Медиа: внешние расширения человека. М.: Гиперборея; Кучково поле, 2007.

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

Birrer A., Just N. (2024) What we know and don't know about deepfakes: An investigation into the state of the research and regulatory landscape. *New Media & Society*. Online first. DOI: 10.1177/14614448241253138.

Borodin K., Kudryavtsev V., Korzh D., Efimenko A., Mkrtchian G., Gorodnichev M., Rogov O. Y. (2024) AASIST3: KAN-Enhanced AASIST Speech Deepfake Detection using SSL Features and Additional Regularization for the ASVspoof 2024 Challenge. *arXiv preprint* arXiv:2408.17352.

Bray S. D., Johnson S. D., Kleinberg B. (2023) Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity* 9 (1): tyad011. DOI: 10.1093/cybsec/tyad011.

Cover R. (2022) Deepfake culture: the emergence of audio-video deception as an object of social anxiety and regulation. *Continuum: Journal of Media & Cultural Studies* 36 (7): 1–13. DOI: 10.1080/10304312.2022.2084039.

Hameleers M., van der Meer T. G. L. A., Dobber T. (2024) They Would Never Say Anything Like This! Reasons to Doubt Political Deepfakes. *European Journal of Communication* 39 (3): 56–70. DOI: 10.1177/02673231231184703.

Hynek N., Gavurova B., Kubak M. (2025) Risks and benefits of artificial intelligence deepfakes: Systematic review and comparison of public attitudes in seven European countries. *Journal of Innovation & Knowledge* 10 (5): 100782. DOI: 10.1016/j.jik.2025.100782.

Köbis N. C., Doležalová B., Soraperra I. (2021) Fooled twice: People cannot detect deepfakes but think they can. *iScience* 24: 103364. DOI: 10.1016/j.isci.2021.103364.

Mai K. T., Bray S., Davies T., Griffin L. D. (2023) Warning: Humans cannot reliably detect speech deepfakes. *PLoS ONE* 18 (8): e0285333. DOI: 10.1371/journal.pone.0285333.

Mirsky Y., Lee W. (2021) The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)* 54 (1): 1–41.

Notes

Dartmouth Guide to Digital Hygiene. *Dartmouth College*. Available at: https://services.dartmouth.edu/TDClient/1806/Portal/KB/ArticleDet?ID=155669 (accessed: 15.06.2025).

Digital Hygiene. *Group-IB*. Available at: https://group-ib.medium.com/digital-hygiene-338f4c60f24c (accessed: 15.06.2025).

Eksperty soobshchili o rezkom roste chisla moshennichestv s ispol'zovaniem dipfeykov [Experts report a sharp increase in fraud involving deepfakes]. *Forbes*. Available at: https://www.forbes.ru/tekhnologii/539953-eksperty-soobsili-o-rezkom-roste-cisla-mosennicestv-s-ispol-zovaniem-dipfejkov (accessed: 14.09.2025).

Good Digital Hygiene Practices You Need to Know. *Vodafone*. Available at: https://www.vodafone.com.au/blog/security/good-digital-hygiene-practices-you-need-to-know (accessed: 15.06.2025)

Roettgers J. (2018) Porn Producers Offer to Help Hollywood Take Down Deepfake Videos. *Variety*. Available at: https://variety.com/2018/digital/news/deepfakes-porn-adult-industry-1202705749/ (accessed: 15.06.2025).

Semantic Forensics. *DARPA*. Available at: https://www.darpa.mil/program/semantic-forensics (дата обращения: 15.06.2025).

Uncovering Deepfakes: Classroom Guide + Discussion Questions. *AI for Education*. Available at: https://www.aiforeducation.io/ai-resources/uncovering-deepfakes (accessed: 15.06.2025).

Voronin M. S. (2022) Bruce Willis and Georges Miloslavsky: How deepfakes are created and used. *Project and Training Laboratory of Economic Journalism, HSE University*. 11 April 2022. Available at: https://economics.hse.ru/ecjourn/news/588939161.html (accessed: 15.06.2025).

References

Birrer A., Just N. (2024) What we know and don't know about deepfakes: An investigation into the state of the research and regulatory landscape. *New Media & Society*. Online first. DOI: 10.1177/14614448241253138.

Borodin K., Kudryavtsev V., Korzh D., Efimenko A., Mkrtchian G., Gorodnichev M., Rogov O. Y. (2024) AASIST3: KAN-Enhanced AASIST Speech Deepfake Detection using SSL Features and Additional Regularization for the ASVspoof 2024 Challenge. *arXiv preprint* arXiv:2408.17352.

Bray S. D., Johnson S. D., Kleinberg B. (2023) Testing human ability to detect ‘deepfake’ images of human faces. *Journal of Cybersecurity* 9 (1): tyad011. DOI: 10.1093/cybsec/tyad011.

Cover R. (2022) Deepfake culture: the emergence of audio-video deception as an object of social anxiety and regulation. *Continuum: Journal of Media & Cultural Studies* 36 (7): 1–13. DOI: 10.1080/10304312.2022.2084039.

Hameleers M., van der Meer T. G. L. A., Dobber T. (2024) They Would Never Say Anything Like This! Reasons to Doubt Political Deepfakes. *European Journal of Communication* 39 (3): 56–70. DOI: 10.1177/02673231231184703.

Hynek N., Gavurova B., Kubak M. (2025) Risks and benefits of artificial intelligence deepfakes: Systematic review and comparison of public attitudes in seven European countries. *Journal of Innovation & Knowledge* 10 (5): 100782. DOI: 10.1016/j.jik.2025.100782.

Köbis N. C., Doležalová B., Soraperra I. (2021) Fooled twice: People cannot detect deepfakes but think they can. *iScience* 24: 103364. DOI: 10.1016/j.isci.2021.103364.

Mai K. T., Bray S., Davies T., Griffin L. D. (2023) Warning: Humans cannot reliably detect speech deepfakes. *PLoS ONE* 18 (8): e0285333. DOI: 10.1371/journal.pone.0285333.

Mirsky Y., Lee W. (2021) The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)* 54 (1): 1–41.

Ashmanov I. S., Kasperskaya N. I. (2021) *Tsifrovaya gigiena* [Digital hygiene]. St. Petersburg: Piter Publ. (In Russian)

Ivanov V. G., Ignatovsky Ya. R. (2020) Deepfakes: perspektivy primeneniya v politike i ugrozy dlya lichnosti i natsional’noy bezopasnosti [Deepfakes: Prospects for political use and threats to individual and national security]. *RUDN Journal of Public Administration* 7 (4): 379–386. (In Russian)

Kiselev A. S. (2021) O neobkhodimosti pravovogo regulirovaniya iskusstvennogo intellekta: dipfeyk kak ugroza natsional’noy bezopasnosti [On the need for legal regulation of artificial intelligence: deepfake as a threat to national security]. *Vestnik Moskovskogo gosudarstvennogo oblastnogo universiteta. Seriya: Yurisprudentsiya* 3: 54–64. (In Russian)

Krasovskaya N. R., Gulyaev A. A. (2020) Tekhnologii manipulyatsii soznaniem pri ispol’zovanii dipfeykov kak instrumenta informatsionnoy voyny v politicheskoy sfere [Technologies of consciousness manipulation using deepfakes as an information warfare tool in politics]. *Vlast’* 28 (4): 93–98. (In Russian)

McLuhan M. (2007) *Understanding Media: The Extensions of Man*. Moscow: Giperboreya; Kuchkovo Pole Publ. (In Russian)

Sukhodaeva T. S. (2022) Formirovanie navykov tsifrovoy gigieny studentov kak sushchestvennyy element vospitatel’noy raboty v vysshey shkole [Formation of students’ digital hygiene skills as an essential element of educational work in higher education]. *Vestnik SGUPS: Gumanitarnye issledovaniya* 2 (13): 100–105. (In Russian)

Поступила в редакцию
15.06.2025