ИССЛЕДОВАТЕЛЬСКАЯ СТАТЬЯ

Искусственный интеллект и дипфейки: вызовы и перспективы*

Елена Александровна Литаш-Сорокина[a]

a Центральный банк Российской Федерации

EDN: HNSWCA

Аннотация: Искусственный интеллект (ИИ) глубоко проник в повседневную жизнь, радикально изменив общественные структуры, экономику и политические процессы. Современные достижения в области ИИ, особенно в области глубокого обучения и генеративного ИИ (GenAI), открывают новые горизонты возможностей, но также создают серьезные вызовы. Статья исследует текущее состояние и перспективы развития ИИ, его влияние на общество и политику, а также возникающие риски и возможности. Особое внимание уделяется феномену дипфейков – видео, созданных с помощью ИИ, которые могут использоваться для дезинформации и манипуляции общественным мнением. В статье также рассматриваются угрозы, связанные с возможным выходом ИИ из-под контроля человека, отмечается необходимость правового регулирования и международного сотрудничества для предотвращения потенциальных угроз. В заключение подчеркивается важность управления развитием ИИ с учетом его преимуществ и недостатков.

Ключевые слова: искусственный интеллект (ИИ), общество, государство, дипфейки, политическая коммуникация, социальные сети, дезинформация, киберпреступность, цифровая грамотность, этические вопросы

Дата поступления статьи в редакцию: 21 марта 2025 года.

Artificial Intelligence and deepfakes: challenges and prospects**

RESEARCH ARTICLE

Elena A. Litash-Sorokina[a]

a The Central Bank of the Russian Federation

Abstract: Artificial Intelligence has become an integral part in everyday life, transforming social structures, the economy, and political processes. The latest advances in AI, particularly deep learning and generative AI (GenAI), bring new opportunities while also posing significant challenges. The article looks at the present and potential prospects of AI development, its impact on society and politics, and new threats and opportunities. The author focuses on the phenomenon of deepfakes – AI-generated videos that can be exploited to propagate misinformation and manipulate public opinion. The article also explores the risks connected with AI potentially getting out of human control, emphasizing the importance of legal regulation and international cooperation in preventing such dangers. In conclusion, the author highlights the significance of monitoring AI development while considering its benefits and drawbacks.

Keywords: artificial intelligence (AI), society, state, deepfakes, political communication, social media, disinformation, cybercrime, digital literacy, ethical issues

Received: March 21, 2025.

* Содержание настоящей статьи отражает личную позицию автора. Содержание и результаты статьи не следует рассматривать, в том числе цитировать в каких-либо изданиях, как официальную позицию Банка России или указание на официальную политику или решения регулятора. Любые ошибки в данном материале являются исключительно авторскими.

** This article's material is based on the author's personal opinions. The content and findings of this article should not be interpreted or quoted in any media as the Bank of Russia's official opinion or as an indicator of the regulator's official policy or decisions. Any mistakes in this material will be the sole responsibility of the author.

Возможно, придется просто попрощаться
с возможностью быстро узнать правду,
с тем, что можно просто быстро загуглить
и узнать, где факт, а где вымысел, –
я думаю, это больше не работает

Сандра Вахтер,
профессор Оксфордского института Интернета,
исследователь правовых и этических
последствий ИИ

Введение

В последние годы значение искусственного интеллекта (ИИ) в общественной жизни резко возросло. Он оказывает существенное влияние на людей, работу корпораций, рынков, правительств и общество в целом, что делает необходимым изучение как его положительных, так и отрицательных аспектов.

ИИ не является новой областью исследований. Работа над машинным обучением ведется с конца 30-х годов прошлого века, хотя сам термин ИИ стал популярным только в 1950-х, на фоне всплеска интереса к обработке естественного языка, машинному обучению и робототехнике. Ограничения вычислительных мощностей привели к истощению финансирования таких проектов и «зиме ИИ» в 1970-х годах. В последующие десятилетия наблюдалось чередование периодов интенсивных исследований и спада интереса к проблеме. На протяжении всего этого времени программное обеспечение ИИ разрабатывалось для конкретных целей, основываясь на четких правилах и параметрах, определяемых человеком[1]. В основе современного ИИ лежит машинное обучение: его системы пополняются примерами входных данных желаемого результата, что позволяет создавать модели, применимые к совершенно новым вводным. Чем к большим объемам информации есть доступ, тем лучше модели учатся и работают. Используемый сейчас подход называется глубоким обучением, поскольку искусственный интеллект выявляет закономерности в огромном количестве наборов данных, характеризующихся разной структурой. По сути, ИИ учится, постоянно переоценивая свои знания, формируя новые связи и расставляя приоритеты на основе новых данных, с которыми он сталкивается. Результаты, особенно в таких областях, как распознавание и генерация изображений и речи, впечатляют, однако процесс работы ИИ остается не всегда ясным даже для экспертов. Современные модели генеративного ИИ (GenAI) обучаются данным из различных областей, без каких-либо ограничений с точки зрения задачи, в отличие от более ранних моделей. И если до 2022 года чат-боты (ассистенты на сайтах компаний или государственных учреждений), к которым все успели привыкнуть, следовали скриптовым ответам и полагались на предопределенные правила взаимодействия с пользователями, современные чат-боты с GenAI, такие как ChatGPT или Google Gemini, могут воспроизводить текст, адаптируясь ко многим темам без ограничений какой-либо логикой, что позволяет полноценно участвовать в диалоге с людьми. Кроме того, такие чат-боты могут создавать не только текст, но и изображения, музыку, компьютерные коды на основе наборов данных, на которых они были обучены.

ИИ проник в нашу жизнь постепенно, через все, что связано с технологиями: от смартфонов и автомобилей до маркетинговых кампаний продаж и систем, обеспечивающих функционирование городов. В результате его прогресс для человека стал почти незаметен. AlphaGo, разработанная компанией DeepMind, использовала ИИ и победила чемпиона мира по игре в го в 2016 году, но быстро стерлась из общественного сознания. GPT-1, модель, представленная в июне 2018 года, стала первой итерацией серии GPT (генеративный, предварительно обученный трансформатор). В ноябре 2022 года был выпущен ChatGPT, который довольно быстро стал вирусным, а уже через четыре месяца OpenAI выпустила новую версию, названную GPT-4, заметно улучшив возможности. В мае 2023 года Google, Microsoft и другие компании анонсировали новые функции на базе GenAI, и уже сегодня новейшие приложения, использующие GenAI, могут выполнять для человека множество рутинных задач по реорганизации и классификации существующих данных и созданию нового контента. Сегодня использование ИИ приводит к трансформации производственных ролей, росту производительности, укреплению позиций среднего класса [Camilleri, 2023], но именно способность писать текст, сочинять музыку и создавать цифровые произведения искусства побуждает пользователей экспериментировать с ИИ самостоятельно. В результате все более широкий круг заинтересованных сторон сталкивается с влиянием GenAI на бизнес и общество, однако не всегда осознает его возможности, которые позволили бы использовать его в полной мере.

Постановка проблемы. «Серые зоны» ИИ

Технологические достижения в области ИИ оказывают влияние на экономическую деятельность, рынки труда и общественные структуры, меняют общество. Помимо новых возможностей осознаются и потенциальные риски, связанные с широкими социальными, этическими и экономическими по-

1 Generative AI – new systems with a long history. https://www.wipo.int/web-publications/patent-landscape-report-generative-artificial-intelligence-genai/en/introduction.html

следствиями для человечества. ИИ-пессимисты меняются местами с ИИ-оптимистами[2]. Достижения ИИ вызывают серьезные опасения бизнеса, правительств, в академических кругах и гражданском обществе, касающиеся риска возможного вреда. ИИ играет все более заметную роль в динамике повседневности, что делает необходимым осознание потенциальных последствий его применения в политической сфере, роли в политической коммуникации. Особое внимание следует уделить феномену дипфейков, которые потенциально ставят под угрозу демократическую стабильность и безопасность. Благодаря прогнозному анализу данных и обработке информации из различных источников ИИ позволяет предвыборным командам более точно идентифицировать целевую аудиторию, персонализировать политические сообщения с учетом конкретных потребностей и предпочтений избирателей, повышая эффективность кампании. С другой стороны, ИИ предлагает инструменты анализа и моделирования, которые позволяют политикам более точно и своевременно оценивать потенциальные последствия своих действий. Способность обрабатывать большие объемы данных и предвидеть последствия политических решений предоставляет значительные преимущества. Присутствие чат-ботов и ИИ-помощников на веб-сайтах политиков, а также использование социальных платформ для взаимодействия с избирателями служат примерами того, как ИИ может облегчить прямое общение между представителями власти и гражданами. Вместе с тем быстрое, незаметное проникновение ИИ в жизнь граждан поднимает важные вопросы о будущем человечества. Генеральный директор OpenAI Сэм Альтман на одном из заседаний Сената США выразил обеспокоенность тем, что модели ИИ могут быть использованы для масштабных кампаний по дезинформации или кибератак, призвав политиков принять меры для его регулирования[3]. Игнорируя эти угрозы, человечество может необратимо потерять контроль над автономными системами ИИ, сделав вмешательство человека неэффективным. Неконтролируемое развитие ИИ, как полагают эксперты, может привести к кризисам с большим числом человеческих жертв и даже вымиранию человечества [Bengio, Hinton, Yao et al. 2024. P. 842].

В рамках этой статьи мы попытаемся, опираясь на актуальные исследования, открытые данные (Google, «Яндекс», Mediascope, Brandanalytics), информацию о случаях распространения дипфейков в реальном мире, исследовать: 1) потенциал воздействия дипфейков на уязвимые группы людей; 2) последствия влияния развития ИИ, используемого злонамеренным образом; 3) возможности снижения негативных эффектов. Цель этой статьи – предложить видение того, как обезопасить человека в стремительно эволюционирующем, как и сами ИИ-технологии, мире.

Распространение поддельной реальности

Технологию дипфейка предложил Ян Гудфеллоу в 2014 году в ответ на просьбу друзей помочь с проектом, который мог бы сам создавать фотографии[4]. Теперь она называется GAN, или «генеративно-состязательная сеть» (подробнее см.: [Goodfellow et al., 2020]). В скором времени после создания она стала популярна среди пользователей соцсетей и стала использоваться людьми для развлечения (функция FaceSwap позволяла, в частности, заменять лицо одного человека на фото или видео на лицо другого), однако, помимо положительных эмоций, причинила немало неудобств запечатленным на них людям. Состязательность – основа технологии GAN, однако это и главная проблема обнаружения дипфейков. С момента разработки GAN качество дипфейков возросло, распространение технологии создало значительные проблемы и возможности для манипуляции цифровыми медиа [최 창 욱, 정 유 미…, 2023].

Всемирная организация интеллектуальной собственности, основываясь на анализе научных и патентных данных, в отчете за 2023–2024 годы представила обзор ситуации в сфере ИИ. Как оказалось, большинство патентов относится к технологии GAN – в период с 2014 по 2023 год опубликованы 9 700 патентных семейств этого типа моделей, из которых 2 400 – только в 2023 году. Патенты, связанные с GAN, также показывают самый сильный рост за последнее десятилетие, только недавно уступив первенство патентам, связанными с диффузионными и большими языковыми моделями (LLM), что объясняется ростом интереса к GenAI и чат-ботам типа ChatGPT[5]. Среди различных режимов ввода и вывода данных GenAI большинство патентов относятся к категории изображений / видео, которые особенно важны для GAN. Аналитики прогнозируют, что технологии GAN на рынке отрас-

2 An A.I. Researcher Takes on Election Deepfakes. https://www.nytimes.com/2024/04/02/technology/an-ai-researcher-takes-on-election-deepfakes.html

3 OpenAI CEO Sam Altman says he's a «little bit scared» of A.I. https://www.cnbc.com/2023/03/20/openai-ceo-sam-altman-says-hes-a-little-bit-scared-of-ai.html

4 The GANfather: The man who's given machines the gift of imagination. https://www.technologyreview.com/2018/02/21/145289/the-ganfather-the-man-whos-given-machines-the-gift-of-imagination/

5 世界知识产权组织报告：中国生成式AI专利申请量全球领先. https://www.163.com/dy/article/J69A4R8V05149FJG.html

ли дипфейков получат наибольшую долю рынка. И если в 2024 году бóльшая его часть принадлежала североамериканским компаниям, то, как ожидается, к 2030 году самые быстрые темпы роста покажет Азиатско-Тихоокеанский регион. Китай, Южная Корея и Япония обладают надежными технологическими экосистемами, позволяющими развивать технологии дипфейка. Такие крупные игроки, как китайские Baidu и Tencent, лидируют в этой области. Кроме того, огромное население региона в сочетании с ростом цифровизации, использованием смартфонов создает оптимальную среду для инноваций на основе ИИ в медиа, рекламе и играх[6].

Китайская компания Tencent Holdings – мировой лидер по количеству опубликованных патентных семейств в режимах GenAI, которые обрабатывают разные типы данных (изображения / видео, текст, речь / голос / музыку), а также другие режимы, – специализируется на ПО для социальных сетей и видеоигр, а также исследованиях в области GenAI. В первом полугодии 2023 года компания объявила о планах по запуску платформы для создания цифровых человеческих образов – Deepfakes-as-a-Service (DFaaS)[7]. Tencent утверждала, что для создания цифровой копии человека высокой четкости сервису требуется всего три минуты живого видео и 100 произнесенных предложений. «Клону» можно будет задавать вопросы, превращая его в своего рода чат-бот, совмещенный с технологией дипфейк. Tencent также планирует создавать «цифровых» врачей, юристов и других специалистов[8].

Другой пример: стартап Synthesia – одна из немногих европейских компаний в сфере ИИ, получивших статус «единорога», – планирует создавать полноразмерные аватары, которые смогут ходить и перемещаться в реальном пространстве. Глава компании утверждает, что технологии Synthesia уже используют 56 % компаний из списка Fortune 1000[9].

Ключевые угрозы дезинформирующих дипфейков

Само слово «дипфейк» представляет собой комбинацию слов «глубокое обучение» и «подделка» и относится в первую очередь к дезинформации [Walczyna, Piotrowski, 2024]. Фейковая информация распространяется в интернет-среде с использованием механизмов социального маркетинга в качестве сатирического контента или контента, призванного повлиять на людей и сформировать необходимую точку зрения. Этим она отличается от обычной ложной информации. Феномен фейковых новостей – явление не новое; кампании по дезинформации противника известны с глубокой древности (один из первых зафиксированных случаев относится к I веку до н.э. [Watson, 2018. P. 94]). Однако более опасным для людей и общества он становится из-за всеобщей цифровизации и развития GenAI. Благодаря использованию социальных сетей в качестве средств коммуникации и сформированному за два десятилетия доверию к ним фейки распространяются с молниеносной скоростью.

С момента изобретения фотографии визуальные средства массовой информации пользуются высоким уровнем доверия со стороны общественности. Общепризнано, что они особенно эффективны в качестве инструмента пропаганды. Использование сфабрикованных изображений в политических целях документируется как минимум с 20-х годов прошлого века. Однако до появления ИИ создание видео в целях политической пропаганды было дорогим и редким явлением: для производства таких материалов были необходимы квалифицированные специалисты и большое количество времени, так как было необходимо менять каждый кадр. Технология манипулирования видео была усовершенствована в Голливуде в 1990-х годах, но была настолько дорогостоящей, что киностудии смогли использовать ее только в нескольких фильмах [Langguth, Pogorelov et al., 2021]. Сейчас ситуация изменилась настолько, что каждый, кто обладает начальными цифровыми навыками, незначительным набором видео- и фотоматериалов, а также доступом к компьютеру, может создавать фейковые видео.

Дипфейки печально известны из-за своего неэтичного и вредоносного применения для достижения экономических, политических и социальных целей. Несмотря на вводимые странами запреты[10], новости об обнаружении дипфейков не покидают первые полосы мировых СМИ: арест Трампа[11], Па-

6 Deepfake AI Market Size, Share, and Growth Analysis, 2024. https://www.researchandmarkets.com/reports/5983205/deepfake-ai-market-size-share-growth-analysis

7 Tencent Cloud announces Deepfakes-as-a-Service for $145. https://www.theregister.com/2023/04/28/tencent_digital_humans/

8 腾讯云公布小样本数智人生产平台，花费千元即可自行制作数字人. https://www.jiemian.com/article/9312569.html

9 The people onscreen are fake. The disinformation is real. https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html

10 The Strengthened Code of Practice on Disinformation. https://regmedia.co.uk/2022/06/16/eu_code_of_practice_2022.pdf; Katyanna Quach, China follows through on plan to ban deepfake tech. https://www.theregister.com/2023/01/10/china_deep_fake/

11 AI-Generated Images of Donald Trump Getting Arrested Foreshadow a Flood of Memes, Fake News. https://www.forbes.com/sites/danidiplacido/2023/03/22/ai-generated-images-of-donald-trump-getting-arrested-foreshadow-a-flood-of-memes-fake-news/

па Римский Франциск в дизайнерском пуховике[12], пропагандистские видео с ведущими новостей[13], распространяемые бот-аккаунтами в Facebook (запрещена в РФ) и X (запрещена в РФ) в рамках информационной кампании[14], и многое, многое другое. Исследователи, журналисты и чиновники выражают обеспокоенность широким использованием ИИ-контента в политической деятельности, активно обсуждая возможные сценарии реализации угроз национальной и международной безопасности.

Если ранее присутствие в Сети дипфейков сводилось преимущественно к фейковым изображениям лиц и видеоматериалам, демонстрирующим якобы совершенные людьми действия и вкладывающие в их уста поддельные цитаты, то уже в самом начале 2023 года стало известно, что персонажей, целиком смоделированных искусственным интеллектом, используют компании, действующие в интересах правительств некоторых стран[15]. Видеоконтент, на котором политики якобы совершают некие действия, либо транслирующий их искаженные взгляды, способен необратимо разрушить их репутацию, серьезно снизить уровень доверия населения, затруднить эффективное выполнение функций власти. В Индии дипфейки стали инструментом давления на женщин-политиков и журналисток. Сообщалось о многочисленных инцидентах, когда их лица использовались в порнографическом контенте, массово распространяемом онлайн с целью очернить женщин-лидеров, существенно подрывая их имидж и снижая шансы на успех [Pawelec, 2022; Gill, 2024].

Дезинформационные кампании, усиленные технологией дипфейка, представляют собой угрозу политической стабильности и эффективному управлению, усложняя процесс принятия решений руководителями государств и международными структурами. Наибольший риск использования дипфейков в политике кроется даже не в самой дезинформации, а в непредсказуемом воздействии на сознание аудитории, усугубляемом низкой осведомленностью общества относительно масштабов и потенциала применения ИИ. Исследования в этой области подтверждают, что дипфейки оказывают влияние на восприятие конкретного политика, оказывая меньшее влияние на восприятие стоящей за ним институции. Когда упомянутая выше Н. Пелоси в дипфейковых видео озвучивала идеи, резко контрастирующие с ее прежними заявлениями, общество расценило перемены как признаки некомпетентности, нестабильности взглядов и разрыва с собственной партией [Hameleers et al., 2024]. Аналогичный эффект был зафиксирован в эксперименте, когда в дипфейковом видео представитель нидерландской партии «Христианско-демократический призыв» якобы пошутил о Распятии Христа, что вызвало негативную реакцию по отношению персонально к нему, но практически не повлияло на восприятие в обществе самой партии. Микротаргетинг усиливал негативные результаты, но лишь в небольших подгруппах наиболее радикально приверженной взглядам партии аудитории [Dobber et al., 2021. P. 75–77]. Эксперты также приходят к выводам, что наибольшее доверие вызывают дипфейки, которые транслируют идеи, близкие к действительным политическим позициям своей жертвы, тогда как сильно искаженная информация, наоборот, снижает доверие. Общая тенденция такова: чем больше человек поддерживал дискредитируемое лицо, с тем меньшей вероятностью он негативно оценит его действия и после демонстрации дипфейка. Одно из объяснений этому может заключаться в том, что когда люди хорошо знакомы с человеком, они могут лучше заметить неестественное поведение или выражение нетипичных для него взглядов. Полученные результаты также свидетельствуют о том, что предшествующая поддержка кандидата или партии имеет бóльшую силу, чем попытки дискредитации [Hameleers et al., 2024]. При этом иногда, как отмечают другие исследователи, тезисы, совершенно не соответствующие реальной идеологии политика, ставшего жертвой дипфейка, парадоксальным образом выглядят достоверно в глазах аудитории [Pennycook, 2021. P. 590].

Возможно, по причине того, что технология дипфейков пока не имеет широкого применения, фейковый контент так и не сыграл заметной роли в политике, о чем изначально беспокоились многие специалисты [Langguth, Pogorelov et al., 2021]. Не подтвердились и опасения, что дипфейки могут быть использованы для значимых манипуляций избирателями, возможных благодаря способности ИИ обрабатывать информацию в режиме реального времени и адаптироваться к индивидуальным предпочтениям и изменению восприятия общественности. Однако это не означает, что этой угрозы не существует, особенно в свете бурного развития ИИ.

12 Fake photos of Pope Francis in a puffer jacket go viral, highlighting the power and peril of AI. https://www.cbsnews.com/news/pope-francis-puffer-jacket-fake-photos-deepfake-power-peril-of-ai

13 Let's play a game: Deepfake news anchor or a real person? https://www.theregister.com/2023/02/11/deepfake_news_anchors/

14 The Disinformation Is Real. https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html

15 Deepfake It till You Make It. https://graphika.com/reports/deepfake-it-till-you-make-it

В политике ИИ применяется с различными целями. Так, политические партии активно используют дипфейки для поддержки своих предвыборных кампаний. Например, Национальный комитет Демократической партии США использовал дипфейк председателя комитета Тома Переса на хакерском съезде DEF CON 2019, чтобы привлечь внимание к дипфейкам как к потенциальной угрозе выборам в США 2020 года[16]. На выборах в Южной Корее 2022 года две основные партии использовали дипфейки своих кандидатов для общения с избирателями[17]. В 2023 году партия «Шведские демократы» опубликовала дипфейк, на котором лидер партии Джимми Окессон произнес речь на арабском языке с целью охватить аудиторию, не говорящую по-шведски[18]. Политики Индии активно используют цифровые копии самих себя, что связано со спецификой и размером их аудитории: в стране, население которой превышает 1,4 млрд человек, 22 официальных языка и тысячи региональных диалектов. Технология позволяет отправлять персонализированные сообщения в отдаленные общины, жители которых и не подозревают, что общаются с цифровым клоном[19]. Распространение дипфейков известных политиков по всему миру осложняет процедуры верификации фактов, создавая атмосферу неопределенности и усугубляя международные конфликты. Чтобы оценить воздействие дипфейков на репутацию политиков в мировом масштабе, важно учитывать специфику восприятия ложной информации в конкретной стране, степень узнаваемости фигур внутри страны и на мировой арене. В руках злоумышленников дипфейки могут стать оружием, способным спровоцировать внутренние или международные кризисы. Потеря человеком контроля над ИИ может привести к непредсказуемым последствиям. Во время вооруженных конфликтов широкое использование дипфейковых видео приводит к тому, что люди все меньше доверяют видеоматериалам, в том числе подлинным. Более того, регулярное распространение манипулятивного контента провоцирует не только дискредитацию конкретных политиков, но и порождает общее недоверие ко всей поступающей информации и визуальному материалу. Это ведет к так называемому «информационному апокалипсису», когда грань между фактами и фикциями стирается окончательно и люди теряют мотивацию отделять правду от лжи [Stover, 2018].

Факторы распространения дипфейков

Распространению дезинформации среди тысяч (если не миллионов) людей одновременно способствуют социальные сети, поскольку их алгоритмы специально создавались для того, чтобы любое взаимодействие помогало контенту охватить все большее количество людей. Соцсети уверенно лидируют в качестве каналов распространения дипфейков, в том числе злонамеренных: доля YouTube составляет 49,0 %, TikTok – 42,8 %, на запрещенные в РФ Instagram и Facebook приходится 37,9 и 36,2 % соответственно[20]. Гневные реакции или поток комментариев, называющих публикацию ложной, только помогают охватить еще большую аудиторию, поскольку алгоритмы отслеживают лишь степень интереса к публикации, не пытаясь верифицировать тот или иной контент. Рейтинги и рекомендации от популярных блогеров расширяют возможности распространения дезинформации. Автовоспроизведение также облегчает случайный, непреднамеренный просмотр: можно случайно столкнуться с дезинформацией, которая в дальнейшем будет использована уже алгоритмами рекомендаций. Невнимательность пользователей в 51,2 % случаев становится причиной обмена дезинформацией[21]. Люди, сами того не понимая, становятся разносчиками ложных сведений, «расшаривая» и комментируя дипфейки. Распространению дипфейков способствует эффект «эхо-камеры»: чем больше человек взаимодействует с определенным контентом, тем больше видит подобные материалы. Рекомендуемый контент сужает круг аналогичных сообщений для чтения, видео для просмотра или групп для присоединения. Это приводит к тому, что некоторые пользователи, до этого нечасто сталкивавшиеся с противоположными их точке зрения убеждениями, в своих взглядах могут стать более радикальными, будучи «заперты в эхо-камере», где подвергаются воздействию все более экстремального контента[22].

Ряд экспертов, которые изучали то, как люди распознают дипфейки, сравнивая их навыки с воз-

16 The Democratic Party deepfaked its own chairman to highlight 2020 concerns. https://edition.cnn.com/2019/08/09/tech/deepfake-tom-perez-dnc-defcon

17 AI and Elections: Lessons from South Korea. https://thediplomat.com/2024/05/ai-and-elections-lessons-from-south-korea/

18 Jimmie Åkesson Talar till Nationen – På Arabiska. https://www.gp.se/politik/jimmie-akesson-talar-till-nationen-pa-arabiska-.3699de36-4dcc-4851-90a0-e0abb01daa9e

19 Indian Voters Are Being Bombarded with Millions of Deepfakes. Political Candidates Approve. https://www.wired.com/story/indian-elections-ai-deepfakes/

20 4 Findings in Our Deepfake and Voice Clone. Customer Report. https://www.pindrop.com/blog/findings-in-our-deepfake-and-voice-clone-consumer-report

21 Why do Americans share so much fake news? One big reason is they aren't paying attention, new research suggests. https://www.media.mit.edu/articles/why-do-americans-share-so-much-fake-news-one-big-reason-is-they-aren-t-paying-attention-new-research-suggests/

22 What is an echo chamber? https://edu.gcfglobal.org/en/digital-media-literacy/what-is-an-echo-chamber/1/

можностями ИИ-детекторов, выявили, что участники исследований в среднем верно идентифицировали только 63,3 % дипфейков. Положительный результат во многом зависел от личности участника эксперимента, процент правильных ответов также сильно разнился в зависимости от качества контента [Birrer, Just, 2024]. При этом 70 % участников опроса, проведенного компанией McAfee, признались, что не смогли с уверенностью отличить сгенерированную ИИ аудиозапись от голоса реального человека[23]. Как отмечают эксперты, люди лучше распознают дипфейки, если предполагаемые демографические характеристики (возраст, пол и этническая принадлежность) изображенного человека соответствуют их собственным [Birrer, Just, 2024]. Среди наиболее уязвимых для манипулирования групп – пожилые люди, молодежь, а также жители села. В другом исследовании способности поколения зумеров оценивать источник информации были названы пугающими, так как показали, что более чем 80 % старшеклассников верят в «спонсируемый контент» и опубликованные в Сети фото[24]. Вместе с тем недостаточность изученности взаимосвязи факторов, определяющих способность человека идентифицировать дипфейки, а также противоречивые данные проведенных исследований говорят о необходимости продолжения изучения вопроса.

Привычный людям способ коммуникации в социальных сетях, с помощью видеоконференций, электронной почты и мессенджеров, делает возможным использование граждан не только как распространителей информации «из уст в уста», но и как источник информации об их поведенческих особенностях, предпочтениях и интересах. В результате становятся возможны такие ситуации, как скандал 2016 года, связанный с крупной британской консалтинговой компанией Cambridge Analytica. Она использовала учетные данные пользователей и опросов для формирования подробных психологических профилей и персонализированных сообщений с целью создания узконаправленной политической рекламы и влияния на мнение избирателей. Обвинения в конечном счете подтвердило и руководство компании[25].

Как отмечалось ранее, политическая дезинформация может привести к социальным потрясениям и потере доверия к ключевым институтам. В современном мире технологий и цифровых коммуникаций сложно отследить первоисточник таких сведений. Люди могут подвергаться воздействию большого количества контента в течение какого-то времени, в том числе в процессе обмена сообщениями, что затрудняет идентификацию отдельных дипфейков как источников воздействия. Как поведение, связанное с обменом информацией, так и ее влияние на общественное мнение может различаться в разных регионах или группах с разным уровнем цифровой грамотности.

Чтобы рассмотреть возможности ИИ выйти за рамки контроля человеком, необходимо проанализировать возможности новейших технологий, которые в последнее время интенсивно развиваются. ИИ научил приложения распознавать текст и изображения, GAN дал машинам что-то похожее на воображение, уникальную способность человеческого разума создавать образы, идеи и концепции, GenAI дает возможность генерировать все более высококачественные данные для обработки и помогает модели реагировать на разные сценарии, новейшие диффузные модели помогают генерировать пиксели[26]. Современные ИИ-технологии все менее зависимы от обучающих их людей и способны к «неконтролируемому обучению»: беспилотный автомобиль может обучаться множеству различных дорожных условий, не покидая гаража, а робот-погрузчик может предвидеть препятствия, с которыми он может столкнуться на загруженном складе, без необходимости его прохождения[27].

Опасность заключается в том, что в ИИ-системах относительно легко заменить полезную цель, например, обобщение отчета, на вредоносную, например, создание дезинформации, путем изменения их инструкций. Интуитивный интерфейс на естественном языке позволяет неспециалистам ставить для ИИ злонамеренные цели, как это недавно было продемонстрировано в случае с GPT-4, которого уговорили дать совет по созданию патогенов пандемического уровня[28]. «Довольно легко превратить такую систему, как ChatGPT, в систему, которая может действовать в Интернете без участия человека, что значительно увеличивает ее потенциальный вред. К счастью, ChatGPT пока недостаточно компетентен, чтобы такая система мог-

23 Beware the Artificial Impostor. https://www.mcafee.com/content/dam/consumer/en-us/resources/cybersecurity/artificial-intelligence/rp-beware-the-artificial-impostor-report.pdf

24 Students Have 'Dismaying' Inability to Tell Fake News from Real, Study Finds. https://www.npr.org/sections/thetwo-way/2016/11/23/503129818/study-finds-students-have-dismaying-inability-to-tell-fake-news-from-real

25 Cambridge Analytica and Facebook: The Scandal and the Fallout So Far. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

26 An AI startup made a hyperrealistic deepfake of me that's so good it's scary. https://www.technologyreview.com/2024/04/25/1091772/new-generative-ai-avatar-deepfake-synthesia

27 The GANfather: The man who's given machines the gift of imagination. https://www.technologyreview.com/2018/02/21/145289/the-ganfather-the-man-whos-given-machines-the-gift-of-imagination

28 AI Safety Newsletter #10. https://newsletter.safe.ai/p/ai-safety-newsletter-10

ла успешно достигать сложных вредоносных целей, но ситуация может измениться в течение нескольких лет», – заявил один из самых авторитетных и цитируемых исследователей ИИ Йошуа Бенжио в своем выступлении на заседании подкомитета Сената США по конфиденциальности, технологиям и законодательству в области надзора за ИИ. Эксперт призвал к вмешательству правительства в регулирование ИИ, а также к инвестициям в исследования для смягчения потенциально катастрофических последствий развития ИИ как технологии, приближающейся к уровню когнитивных способностей человека [Bengio, 2024][29].

Концепция Маршала Маклюэна, основанная на идее о том, что средства массовой информации являются «естественными ресурсами» [McLuhan, 1964. P. 21], делает борьбу за возможность распространения информации одной из важнейших задач настоящего. Ведь медиа не просто передают информацию, они транслируют обществу свое представление о мире. Этика, как один из полюсов социальной и моральной экологии, ставит исследование этических проблем цифровой медиакоммуникации на первое место. Все это делает актуальным изучение влияния ИИ на информационную среду.

Ретроспектива кейсов и экспериментов

Хотя дезинформация не является чем-то новым, тема получила широкий резонанс в 2016 году на фоне президентских выборов в США[30] и референдума по Brexit в Великобритании, в ходе которого полностью сфабрикованные истории (представляемые как истинные) получили широкое распространение через социальные сети. Хорошо известны примеры дипфейков с участием американских президентов Барака Обамы и Дональда Трампа, других известных политиков. В 2018 году Фламандская социалистическая партия тиражировала дипфейк Трампа, высмеивающего климатическую политику Бельгии[31]. В 2019 году созданный во Франции дипфейк Трампа заявил, что СПИД искоренен[32]. Нэнси Пелоси (в то время – спикер Палаты представителей Конгресса США) в мае 2019 года стала объектом насмешек из-за распространения видео, сфабрикованного ИИ, где она выглядит пьяной во время публичного выступления. Ролик быстро стал вирусным, и Дональд Трамп тоже поделился его измененной версией в социальных сетях[33]. В 2020 году активисты движения Extinction Rebellion создали дипфейк тогдашнего премьер-министра Бельгии, где Софи Вильмес выступила с речью о предполагаемой связи между COVID-19 и изменением климата[34].

Разоблачающий аудиоклип в преддверии выборов в Словакии. За несколько дней до важнейших национальных выборов в Словакии в социальных сетях был широко распространен разоблачающий аудиоклип, содержащий голосовой клон лидера словацкой Прогрессивной партии Михала Шимечки, в котором он описывал схему фальсификации результатов голосования, в том числе путем подкупа цыганского населения страны. Публикация вызвала резко негативную реакцию политического сообщества Словакии, некоторые представители которого поверили в подлинность видео[35].

Сфабрикованные ролики с участием Риши Сунака. В период с 8 декабря 2023 года по 8 января 2024 года в соцсетях, принадлежащих Meta (запрещена на территории РФ) было выявлено более 140 дипфейковых видео с премьер-министром Великобритании Риши Сунаком. Охват кампании составил более 400 000 человек, в том числе с явным нарушением некоторых рекламных политик Meta (запрещена в РФ). В одном из роликов диктор новостей BBC, похожий на журналистку Сару Кэмпбелл, произносит: «Люди возмущаются, узнав, что уже несколько месяцев подряд Риши Сунак тайно зарабатывает на проекте, расходуя колоссальные суммы, которые изначально предназначались для простых граждан». Далее она утверждает, что Илон Маск запустил приложение, способное автономно проводить биржевые и рыночные операции, и показывает сфабрикованные кадры с участием британского премьера, на которых Сунак дает «комментарии», подтверждая надежность проекта и выражая благодарность Маску за «выбор страны в качестве первой, где будет работать это приложение». Отмечается, что это была первая скоординированная кампания с использованием фальсифицированного видео политического деятеля Великобритании такого уровня[36].

29 Более взвешенный и прагматичный подход предложен в статье: [Johnson, Menzies, 2024. P. 137].

30 Fake News Is A Real Problem. https://www.statista.com/chart/6795/fake-news-is-a-real-problem/

31 A Belgian Political Party Is Circulating A Trump Deepfake Video. https://www.buzzfeednews.com/article/janelytvynenko/a-belgian-political-party-just-published-a-deepfake-video

32 https://www.brut.media/us/videos/us/society/why-an-aids-charity-made-a-deepfake-of-trump

33 https://www.theguardian.com/technology/2019/may/24/facebook-leaves-fake-nancy-pelosi-video-on-site

34 https://www.brusselstimes.com/106320/xr-belgium-posts-deepfake-of-belgian-premier-linking-covid-19-with-climate-crisis

35 Údajná nahrávka telefonátu predsedu PS a novinárky Denníka N vykazuje podľa expertov početné známky manipulácie. https://fakty.afp.com/doc.afp.com.33WY9LF

36 Slew of deepfake video adverts of Sunak on Facebook raises alarm over AI risk to election. https://www.theguardian.com/technology/2024/jan/12/deepfake-video-adverts-sunak-facebook-alarm-ai-risk-election

Дискредитация главы правительства Сингапура в преддверии выборов. Политики Сингапура также становились жертвами дипфейков. Совсем недавно в соцсетях распространялись созданные ИИ видеоролики с участием главы правительства Ли Сянь Лунга, комментирующего международные вопросы. Ли вскоре опроверг их подлинность, отметив, что эти видео опасны и вредны для национальных интересов[37]. В другом запущенном фейковом видео премьер-министр просит зрителей подписаться на инвестиционный продукт, который, как утверждается, обеспечивает гарантированную прибыль. Ранее, в декабре 2023 года, он уже предупреждал общественность о распространении дипфейка, в котором он выступал в поддержку одной из криптовалютных схем[38].

Более 4,8 млн просмотров минутного видео с участием Барака Обамы. На видео бывший президент США Барак Обама, сидящий на фоне американского флага, говорит с неким зрителем, используя нецензурную лексику в отношении своего преемника Дональда Трампа. Дипфейк был создан Джорданом Пилом, который выдавал себя за Обаму. Пил создал видео, чтобы продемонстрировать опасность сфабрикованного аудио- и видеоконтента, изображающего людей, говорящих или делающих то, чего они на самом деле никогда не говорили и не делали[39].

Международный опыт правового регулирования

В настоящее время не существует международного документа, регулирующего технологию дипфейков. Растущее понимание политиками необходимости таких норм привело к тому, что некоторыми странами и наднациональными объединениями были приняты важные меры.

Рост количества нормативных актов, связанных с ИИ, пропорционален степени развития технологии и проникновения дипфейков в социальные медиа, росту количества скандалов с участием публичных лиц. Так, например, на фоне манипулирования данными и использования дипфейков на выборах в США к 2023 году было предложено 269 законопроектов и принято 25 против одного в 2016. Растет количество органов, регулирующих ИИ[40]. Количество упоминаний ИИ в законодательных актах по всему миру почти удвоилось: с 1 247 в 2022 году до 2 175 в 2023 году. В 2023 году ИИ упоминался в законодательных актах 49 стран[41]. Государства Европейского союза достигли соглашения по условиям знакового Закона об ИИ в 2024 году[42].

Ключевая проблема такого законодательства состоит в том, что социальные сети в большинстве юрисдикций вправе самостоятельно решать вопрос о размещении на их платформах того или иного контента. Однако, ссылаясь на последний случай, связанный с публикацией дипфейковых изображений певицы Тейлор Свифт, власти США признали ответственность социальных сетей за распространение дезинформации и интимных изображений реальных людей без их согласия[43]. Основная проблема заключается в необходимости разработки специального законодательства или правил, которые защищают право на личность и вместе с тем не вступают в противоречие со свободой слова и выражения мнений. Регулирование политических дипфейков представляет собой серьезную проблему. В большинстве стран любой законодательный акт, направленный на запрет использования дипфейков политиками, вступает в противоречие с правом на свободу выражения. Чтобы избежать этого, придется провести четкое различие между дипфейками, созданными с художественными целями, и дипфейками, используемыми со злым умыслом. Для уголовного преследования людей, использующих дипфейки в злонамеренных целях, должны быть представлены доказательства намерений нанести вред. Дипфейки, созданные для развлечения или с сатирическими целями, чаще всего не имеют целей нанесения вреда [Ramluckan, 2024. P. 284].

Между тем в США нет федеральных законов, призванных бороться с потенциальными угрозами технологии дипфейка. В США в конце 2019 года Конгресс одобрил Закон о полномочиях национальной обороны, который обязывает директора Национальной разведки предоставлять отчеты об использовании дипфейков иностранными правительствами и их потенциальном влиянии на нацио-

37 SM Lee warns public against deepfake videos of him commenting on international matters. https://www.straitstimes.com/singapore/sm-lee-warns-public-against-deepfake-videos-of-him-commenting-on-international-matters

38 SM Lee warns that video of him promoting investment scam on social media is a deepfake. https://www.straitstimes.com/singapore/sm-lee-warns-that-video-of-him-promoting-an-investment-scam-on-social-media-is-a-deepfake

39 How Faking Videos Became Easy – And Why That's So Scary. https://fortune.com/2018/09/11/deep-fakes-obama-video/

40 THE AI INDEX REPORT. https://aiindex.stanford.edu/report/#individual-chapters

41 The number of AI regulations in the United States sharply increases. https://hai.stanford.edu/ai-index/2024-ai-index-report/policy-and-governance

42 Tech World's first major law for artificial intelligence gets final EU green light. https://www.cnbc.com/2024/05/21/worlds-first-major-law-for-artificial-intelligence-gets-final-eu-green-light.html

43 https://www.congress.gov/bill/118th-congress/senate-bill/3696/text

нальную безопасность[44]. В отдельных штатах США также приняты законы, касающиеся технологии дипфейков. В Техасе и Калифорнии с 2019 года запрещено использование дипфейков, которые могут повлиять на выборы. Законодательство Калифорнии, Джорджии и Вирджинии запрещает создание и распространение дипфейковой порнографии без согласия с 2019 года [Weiner, 2024. P. 780–783]. В 2020 году штат Нью-Йорк разрешил подавать в суд на распространителей дипфейков[45]. И хотя штаты по-разному подходят к праву на публичность, обычно истец должен доказать, что его личность имеет некоторую узнаваемую коммерческую ценность и что ответчик без разрешения истца использовал личность последнего в коммерческих целях[46].

В Китае в 2023 году было принято «Положение об управлении глубоким синтезом информационных услуг в Интернете»[47]. Документ запрещает создание дипфейков без согласия пользователя и требует, чтобы в видео было указано, что контент был создан с использованием ИИ. В настоящее время Китай является лишь одной из немногих стран, которые вводят строгий запрет на использование дипфейков с участием человека. Положение преследует две ключевые цели: идти в ногу с быстрым развитием технологий и минимизировать их злонамеренное использование.

Закон Европейского союза (ЕС) об искусственном интеллекте квалифицирует ИИ, который может влиять на выборы, референдумы или индивидуальное избирательное поведение, как «высокорискованный». Документ обязывает поставщиков инструментов ИИ общего назначения помечать созданный искусственным интеллектом контент и выявлять манипуляции.

В Сингапуре, где, как уже упоминалось выше, дипфейки с участием первых лиц страны получили большой резонанс, обсуждается временный запрет на использование дипфейков во время выборов. Законодатели ссылаются на опыт Южной Кореи, которая ранее ввела запрет на политический контент, созданный ИИ, за 90 дней до выборов. Нарушителю грозит срок до семи лет или штраф в размере до 50 млн вон (38 000 долларов)[48]. Однако различия в организации выборов делают невозможным простой перенос успешных мер одной страны на опыт другой. Так, в случае с Сингапуром пример Южной Кореи может оказаться неприменимым из-за более короткого предвыборного этапа[49]. В Сингапуре уже действует ряд законов для борьбы с дезинформацией в Интернете, в том числе Закон о защите от лжи и манипуляций в Интернете, который обязывает тех, кто распространяет ложные сведения, опубликовать исправленную версию или прекратить ее распространение вне зависимости от способа и технологии ее создания[50].

Синтетический контент в России: перспективы и вызовы

Международный опыт демонстрирует недостаточную изученность вопросов влияния дипфейков на жизнь общества. Риск-факторы, упомянутые выше, способствуют непредсказуемости эффектов использования дипфейков. Все упомянутые выше тенденции актуальны и для России.

Доверие. Исследования показывают значительный рост популярности и распространения ИИ среди российских пользователей, особенно среди молодежи и образованных городских жителей, которые видят в ИИ полезный и безобидный инструмент[51]. Примерно 28 % россиян уверены, что использование ИИ повысит качество образования[52], и этот тренд также поддерживается организациями, которые используют синтетический контент для модернизации подходов к организации, улучшения мотивации персонала и снижения финансовых затрат на подготовку образовательных материалов. Так, об успешном использовании цифровых персонажей в обучении сообщали такие компании, как МТС[53] и EMP (E-Promo Group)[54].

Доступность. ИИ в России становится частью нового профессионального стандарта, стимулируя

44 First Federal Legislation on Deepfakes Signed into Law. https://www.wilmerhale.com/en/insights/client-alerts/20191223-first-federal-legislation-on-deepfakes-signed-into-law

45 New York's Right to Publicity and Deepfakes Law Breaks New Ground. https://www.jdsupra.com/legalnews/new-york-s-right-to-publicity-and-98428/

46 The Legal Issues Surrounding Deepfakes. https://www.honigman.com/the-matrix/the-legal-issues-surrounding-deepfakes

47 https://xn--h1aax.xn--p1ai/news/razvitie-zakonodatelstva-knr-v-oblasti-tsifrovoy-ekonomiki/?ysclid=lu61e572xo553621939

48 90-day ban on deepfake political ads passes parliamentary special committee. https://en.yna.co.kr/view/AEN20231205006400315

49 Temporary deepfake ban discussed as way to tackle AI falsehoods during Singapore election. https://www.straitstimes.com/singapore/temporary-deepfake-ban-discussed-as-way-to-tackle-ai-falsehoods-during-singapore-elections

50 Fake news law passed after 2 days of debate. https://www.straitstimes.com/politics/parliament-fake-news-law-passed-after-2-days-of-debate

51 ИИ: ваш новый лучший друг? https://wciom.ru/analytical-reviews/analiticheskii-obzor/ii-vash-novyi-luchshii-drug

52 Пределы доверия: естественный интеллект – об искусственном. https://wciom.ru/analytical-reviews/analiticheskii-obzor/predely-doverija-estestvennyi-intellekt-ob-iskusstvennom

53 МТС будет обучать сотрудников с помощью цифровых аватаров. https://moskva.mts.ru/about/mts-dlya-obshhestva/novosti-i-otcheti/novosti/2024-11-14/mts-budet-obuchat-sotrudnikov-s-pomoshhyu-cifrovyh-avatarov

54 В российской компании сделали цифровые аватары преподавателей для обучения сотрудников. https://skillbox.ru/media/education/v-rossiyskoy-kompanii-sdelali-cifrovye-avatary-prepodavateley-dlya-obucheniya-sotrudnikov/

граждан к самостоятельному использованию технологий: 52 % осваивает работу с нейросетями самостоятельно, через видео, статьи и блоги, а 42 % просто начали использовать их без какого-либо обучения[55]. С ростом доступности технологии производства дипфейков[56] россияне все активнее используют нейросети для создания контента – от генерации музыки до визуальных эффектов. Популярность приложений для создания изображений или разработки сценариев продолжает расти[57]. Помимо самостоятельного создания дипфейков, синтетический контент можно заказать у специалистов или приобрести на цифровых маркетплейсах (например, Genies)[58]. Для разработки аудио и визуального представления можно использовать существующие современные модели от Voice cloning, OpenAI, ElevenLabs, Microsoft и других ведущих разработчиков, благодаря которым доступно создание неотличимого от настоящего человека двойника на основе фото, образцов видео с эмоциями и речью, способного вести диалог как в записи, так и в режиме онлайн (во время звонка или видеозвонка)[59].

Выгода. Приобретающая все большую популярность в интернет-пространстве[60] технология дипфейков обещает россиянам и российским компаниям, институциональным структурам многочисленные преимущества: экономию времени и денег[61]; адаптацию под разные аудитории, задачи и ситуации; привлечение аудитории, увеличение охватов; улучшение взаимодействия с клиентами (в том числе индивидуальный подход) и т.д. Общий тренд запросов, с которыми пользователи Рунета обращаются к поисковым платформам (Google и «Яндекс»), демонстрирует заинтересованность в использовании технологий дипфейка. Следует отметить, что значительная доля запросов связана именно с их созданием и распространением; вопросы идентификации и защиты от поддельных изображений и видео хоть и остаются значимыми для россиян аспектами, но на данный момент второстепенны. Количество запросов, связанных с созданием дипфейков, на июнь 2025 года («создать дипфейк»: «Яндекс» – порядка 28 тыс. запросов за месяц, Google – около 16 тыс. запросов за этот же период) превышает количество вопросов о распознавании («Яндекс» – приблизительно 22 тыс. запросов за месяц, Google – чуть больше 7,4 тыс.). При этом число запросов об отличии фейковых изображений или видео от подлинных выглядит сопоставимым («распознать дипфейк», «отличить дипфейк»: «Яндекс» – примерно 25 тыс. запросов за один месяц, Google – свыше 50,9 тыс. запросов). Наибольший объем выдачи по запросам приходится на платформы общего назначения («Яндекс» и Google), тогда как в научных публикациях (Google.Scholar) технология упоминается значительно реже.

Распространение недостоверной информации. В последнее время для распространения недостоверных сведений все чаще используются ИИ-технологии, их жертвами чаще всего становятся пожилые люди и молодежь[62]. Как показывают опросы, 60 % граждан сталкиваются с фейковыми новостями минимум раз в неделю, 66 % признаются, что хотя бы раз верили дезинформации. Отмечается постепенное (на 11 % по сравнению с 2024 годом) увеличение количества недостоверной информации, связанной с социально-экономическими трендами; граждане стали чаще сталкиваться с дезинформацией об уровне благосостояния населения и фейками, связанными с политической сферой. 75 % опрошенных видят в недостоверной информации угрозу. 52 % сталкивались с зарубежными информационными вбросами, столько же людей перепроверяет иностранные фейки. При этом, опираясь на информацию о динамике интереса к дипфейкам, связанным со специальной военной операцией, можно отметить, что при длительном массированном распространении информации, не подтверждающейся какими-либо фактами, пользователи Сети теряют к ней интерес[63].

Правовое регулирование. Несмотря на положительные аспекты и широкое признание преимуществ ИИ, так же как и во всем мире, остаются нерешенными вопросы безопасности и этики, связанные с его использованием. Так, если создание дипфейка само по себе законно, вызванное им нарушение интеллектуальных прав и личных нематериальных благ (честь, достоинство, репутация, пра-

55 Треть россиян считает ИИ-навыки новым стандартом на рынке труда – как знание ПК в 2000-х. https://www.cnews.ru/news/line/2025-06-25_tret_rossiyan_schitaet_ii-navyki

56 Цифровые аватары. https://av3.studio/sozdanie-cifrovogo-avatara

57 Россияне и нейросети: как ИИ облегчил жизнь людей за год. https://www.cnews.ru/news/line/2025-02-24_rossiyane_i_nejroseti_kak

58 Интеграция цифровых аватаров в бизнес как конкурентное преимущество. https://habr.com/ru/companies/gaz-is/articles/910802/

59 AI-аватар для бизнеса. https://wone-it.ru/services/ai-avatar-dlya-biznesa/

60 Цифровой аватар. Тренды, перспективы, реальные кейсы. https://nanosemantics.ai/webinars/czifrovoj-3-d-avatar-trendy-perspektivy-realnye-kejsy

61 Narrators Production запустила цифровых аватаров людей для бизнеса. https://companies.rbc.ru/news/lUhUZ7rBLw/narrators-productions-zapustila-tsifrovyih-avatarov-lyudej-dlya-biznesa/

62 Маркировка будет отличать оригинальный видео- и аудиоконтент от созданного ИИ. https://tass.ru/ekonomika/24374151

63 В РФ за I квартал 2025 года выявили более 60 уникальных дипфейков. https://tass.ru/obschestvo/23621185

во на имя, изображение и частную жизнь) преступно. Преступным может быть и способ получения материалов для его создания. Меры регулирования дипфейков, как и судебная практика по связанным с ним делам, только формируются. В законодательстве понятие «дипфейк» пока не закреплено, ожидается, что внесенный в сентябре 2024 года в Госдуму законопроект[64] должен это исправить, дополнив Уголовный кодекс наказанием за использование дипфейков в противоправных целях. На данный момент российское законодательство гарантирует защиту изображения человека: его нельзя использовать без согласия, независимо от того, кто и как его разместил (ст. 152.1 ГК). При этом голос человека, также используемый при создании дипфейков, не считается объектом интеллектуальной собственности или нематериальным благом. Чтобы восполнить этот пробел, на рассмотрение Госдумы внесен законопроект о внесении изменений в Гражданский кодекс Российской Федерации, предлагающий осуществлять охрану голоса как объекта личных неимущественных прав гражданина по аналогии с его изображением[65]. Анонсирован запуск новой государственной системы противодействия онлайн-мошенникам уже в 2026 году. Как ожидается, в нее войдет уже действующий сервис «Антифишинг» и новые меры, включая комплексные ИТ-решения[66]. По мнению юристов, гипотетически уже сегодня удаления контента, выпуска опровержения, возмещения убытков и компенсации морального вреда добиться все-таки можно, но для этого понадобится доказать совокупность трех обстоятельств: распространение нарушителем сведений о заявителе; порочащий характер этих сведений; несоответствие сведений действительности[67]. Перспективным решением проблемы может стать маркировка дипфейков при их создании и размещении на зарубежных социальных площадках. Вопрос, однако, пока находится в статусе обсуждения.

Наряду с развитием системы правового регулирования следует отметить положительные тенденции в России: 1) рост осведомленности в области кибермошенничества, поддерживаемый государственными организациями и корпорациями с помощью регулярного обучения сотрудников и информирования клиентов[68]; 2) стремление граждан к самообучению в области ИИ, которое повышает уровень цифровой грамотности и осведомленности о возможностях и угрозах; 3) реализация национальных программ цифровизации («Цифровая экономика» и «Экономика данных»)[69], призванных обеспечить тотальный и равный доступ к цифровым возможностям, информации и медиапространству вне зависимости от места проживания. Вместе с тем для минимизации вреда в ситуации непредсказуемости эффектов необходимо помнить не только о важности формирования осведомленности, но и об этическом аспекте проблемы, для чего необходимо осуществлять регулярный мониторинг, исследовать эффекты влияния ИИ, актуальный уровень ИИ-зрелости аудитории.

Необходимые меры адаптации общества

Дипфейки привлекают интерес общественности по всему миру, так как имеют широкий спектр применения и показывают эффективность в различных отраслях. Прогноз развития рынка дипфейков на ближайшие шесть лет показывает растущий спрос на синтетический контент и медиа. В отсутствие сформированного этического и правового порядка в обществе, одновременно с ростом объемов доступных в Интернете данных как о простых гражданах, так и об общественных деятелях, дипфейки зачастую используются, чтобы атаковать конкретных людей. Безопасность граждан, международная безопасность становятся уязвимыми перед лицом злонамеренного использования дипфейков [Romero Moreno, 2024. P. 299].

Эмпирические данные о распространенности дипфейков практически отсутствуют, нельзя сделать четких выводов о масштабах явления дипфейка и его конкретных проявлениях [Birrer, Just, 2024]. Путаница, согласно теории Жана Бодрийяра, между реальной и поддельной реальностью искажает понимание происходящего [Baudrillard, 2006. P. 454]. Негативное влияние современных цифровых манипуляций на общественное доверие и целостность визуальной информации может дестабилизировать восприятие политических лидеров и управление международными конфликтами. Рост использования ИИ в политике поднимает вопрос о необходимости исследований в области дипфейков, соответствующих государственных мер и повышенного внимания общественности к развивающейся ситуации. Общественные деятели, журналисты, публичные личности со всего мира ведут работу по информированию граждан о том, как работает технология дипфейков и какую угро-

64 Законопроект «О внесении изменений в Уголовный кодекс Российской Федерации» № 718538-8. https://sozd.duma.gov.ru/bill/718538-8

65 Законопроект «О внесении изменений в часть первую Гражданского кодекса Российской Федерации» № 718834-8. https://sozd.duma.gov.ru/bill/718834-8#bh_note

66 RIGF 2025: кибербезопасность и борьба с цифровой дезинформацией. https://rigf.ru/press/rigf-2025-kiberbezopasnost-i-borba-s-tsifrovoy-dezinformatsiey/

67 Цифровой обман: какие права нарушают дипфейки и как за это можно наказать. https://pravo.ru/story/249733/

68 Get Experts, «Навыки будущего». https://getexperts.ru

69 Национальные проекты. https://национальныепроекты.рф/

зу они могут представлять, например, во время выборов[70]. Создаются некоммерческие организации, такие как TrueMedia.org[71], которая предлагает общедоступные инструменты по выявлению дипфейков, и CivAI[72], разрабатывающая бесплатные и общедоступные программы для экспериментов по созданию дипфейков с целью осознания возможностей ИИ.

Вместе с этим, чтобы обеспечить рост информированности населения в области синтетических медиа и дипфейков, снизить негативные эффекты дипфейков, необходимо исследовать их распространение и влияние на общество. Сейчас, когда информации много, но зачастую она недостоверна, а рост использования ИИ в сфере коммуникаций, особенно в сфере политики, сталкивает нас с проблемами, связанными с проверкой источников, манипулированием информацией, еще более значимым становится развитие навыков критического мышления, которые позволят гражданам ориентироваться в среде симулякров и симуляций, сохранить доверие к институтам, источникам информации и реальности. Наступление эпохи ИИ требует формирования эффективных стратегий для обеспечения прозрачного политического дискурса, тщательного анализа способности искусственного интеллекта формировать общественное мнение и влиять на процесс политической коммуникации, с одновременным созданием культуры ответственного производства контента и его распространения. Реализация мер защиты цифровой социальной среды, проверка распространяемых фактов становятся важнейшими задачами в области коммуникации.

Необходимы разработка и продвижение на государственном уровне не только мер законодательного регулирования и финансирования исследований противодействия технологии дипфейка, но и программ медиаграмотности, направленных на повышение устойчивости к ним и укрепление доверия к надежной и фактически точной информации. Они должны обучать тому, как злоумышленники могут создавать дипфейки разного уровня правдоподобности. В свете вышесказанного радикальным, но закономерным шагом видится решение, принятое в ряде стран, – изучение возможностей ИИ в качестве предмета в детских садах и школах[73].

Заключение

Хотя технология дипфейка имеет как положительные, так и отрицательные аспекты, она сталкивается в основном с негативным восприятием, поскольку внимание СМИ сосредоточено главным образом на инцидентах, связанных с ИИ, и возникающих проблемах. Технология дипфейков справедливо вызывает огромное количество вопросов и опасений, однако следует помнить и о ее преимуществах. Например, ИИ-технологии все чаще используются в медицине, образовании и других социально значимых областях.

Вместе с тем не стоит забывать и о потенциальной угрозе, которая представляет собой эта технология. Потеря человечеством контроля над ИИ чревата далеко идущими социальными последствиями. Реализации такого сценария способствует и низкая осведомленность населения, и то, что существующие методы обнаружения дипфейков также работают небезупречно.

Чтобы снизить негативное влияние дипфейков, поставить их на службу обществу, минимизировать риски их использования в злонамеренных целях, обеспечить рост цифровой осведомленности граждан в этой области, требуется реализация комплекса мер, направленного на формирование культуры цифрового потребления, включающего: 1) проведение междисциплинарных исследований; 2) актуализацию программ образования; 3) развитие законодательного регулирования; 4) международный обмен опытом для выработки совместных решений.

70 Can you spot the deepfake? How AI is threatening elections. https://www.youtube.com/watch?v=B4jNttRvbpU

71 Identifying Political Deepfakes in Social Media using AI. https://www.truemedia.org/

72 Building public understanding of AI capabilities and dangers. https://www.civai.org/

73 Ministry of Education introduces AI curriculum in public schools starting from 2025–2026 academic year. https://www.emirates247.com/uae/ministry-of-education-introduces-ai-curriculum-in-public-schools-starting-from-2025-2026-academic-year-2025-05-04-1.739067

Литература

Baudrillard J. The precession of simulacra // Media and cultural studies / Ed. by M.G. Durham, D.M. Kellner. Hoboken: Blackwell Publishing, 2006. P. 453–481. In English

Bengio Y. Government interventions to avert future catastrophic AI risks. *Harvard Data Science Review*. 2024. Special Issue 5. https://doi.org/10.1162/99608f92.d949f941. In English

Bengio Y., Hinton G., Yao A. et al. Managing extreme AI risks amid rapid progress. *Science*. 2024. Vol. 384. Issue 6698. P. 842–845. https://doi.org/10.1126/science.adn011. In English

Birrer A., Just N. What we know and don't know about deepfakes: An investigation into the state of the research and regulatory landscape. *New Media & Society.* 2024. Art. No. 14614448241253138. https://doi.org/10.1177/14614448241253138. In English

Camilleri M.A. Artificial intelligence governance: Ethical considerations and implications for social responsibility. *Expert Systems*. 2024. Vol. 41. Issue 7. Art. No. e13406. https://doi.org/10.1111/exsy.13406. In English

Dobber T., Metoui N., Trilling D., Helberger N., de Vreese C. Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press / Politics*. 2020. Vol. 26. Issue 1.

P. 69–91. https://doi.org/10.1177/1940161220944364. In English

Gill A.N. Ethical implications of AI in shaping online discussions (deepfakes, bots). *Qlantic Journal of Social Sciences and Humanities.* 2024. Vol. 5. No. 4. P. 312–322. https://doi.org/10.55737/qjssh.v-iv(CP).24286. In English

Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair Sh., Courville A., Bengio Y. Generative adversarial networks. *Communications of the ACM.* 2020. Vol. 63. No. 11. P. 139–144. https://doi.org/10.1145/3422622. In English

Hameleers M., van der Meer T.G.L.A., Dobber T. Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreign political deepfakes. *Computers in Human Behavior.* 2024. Vol. 152. Art. No. 108096. https://doi.org/10.1016/j.chb.2023.108096. In English

Johnson B., Menzies T. AI over-hype: a dangerous threat (and how to fix it). *IEEE Software.* Vol. 41. No. 6. P. 131–138. https://doi.org/10.1109/ms.2024.3439138. In English

Langguth J., Pogorelov K., Brenner S., Filkuková P., Schroeder D.Th. Don't trust your eyes: image manipulation in the age of deepfakes. *Frontiers in Communication.* 2021. Vol. 6. Art. No. 632317. https://doi.org/10.3389/fcomm.2021.632317. In English

McLuhan M. Understanding Media: The extensions of man. New York; London: McGraw Hill, 1964.

Pawelec M. Deepfakes and democracy (theory): how synthetic audio-visual media for disinformation and hate speech threaten core democratic functions. *Digital Society.* 2022. Vol. 1. Art. No. 19. https://doi.org/10.1007/s44206-022-00010-6. In English

Pennycook G., Epstein Z., Mosleh M., Arechar A.A., Eckles D., Rand D.G. Shifting attention to accuracy can reduce misinformation online. *Nature*. 2021. Vol. 592. Issue 7855. P. 590–595. https://doi.org/10.1038/s41586-021-03344-2. In English

Ramluckan T. Deepfakes: the legal implications // Proceedings of the 19th International Conference on Cyber Warfare and Security (ICCWS 2024) / Ed. by J. du Toit, B. van Niekerk. Reading: Academic Conferences International Limited, 2024. P. 282–288. https://doi.org/10.34190/iccws.19.1.2099. In English

Romero Moreno F. Generative AI and deepfakes: a human rights approach to tackling harmful content. *International Review of Law, Computers & Technology*. 2024. Vol. 38. Issue 3. P. 297–326. https://doi.org/10.1080/13600869.2024.2324540. In English

Stover D. Garlin Gilchrist: Fighting fake news and the information apocalypse. *Bulletin of the Atomic Scientists*. 2018. Vol. 74. Issue 4. P. 283–288. https://doi.org/10.1080/00963402.2018.1486618. In English

Walczyna T., Piotrowski Z. Fast fake: easy-to-train face swap model. *Applied Sciences.* 2024. Vol. 14. Issue 5. Art. No. 2149. https://doi.org/10.3390/app14052149. In English

Watson C.A. Information literacy in a fake/false news world: an overview of the characteristics of fake news and its historical development. *International Journal of Legal Information.* 2018. Vol. 46. No. 2. P. 93–96. https://doi.org/10.1017/jli.2018.25. In English

Weiner M.D. Destined to Deceive: The Need to Regulate Deepfakes with a Foreseeable Harm Standard. *Michigan Law Review.* 2024. Vol. 122. Issue 4. P. 771–804. https://doi.org/10.36644/mlr.122.4.destined. In English

최창욱, 정유미. 국내외 언론이 바라본 딥페이크 기술. 디지털콘텐츠학회 논문지. 2022. Vol. 23. No. 5. P. 893–904.

References

Changuk Choi, Yumi Jung. How media deal with deepfake technology: comparison of domestic and overseas media coverage. *Journal of Digital Contents Society.* 2022. Vol. 23. No. 5. P. 893–904. https://doi.org/10.9728/dcs.2022.23.5.893. In Korean

ИНФОРМАЦИЯ ОБ АВТОРЕ:

Елена Александровна Литаш-Сорокина, начальник Центра операций с цифровыми рублями Департамента национальной платежной системы
Центральный банк Российской Федерации (Российская Федерация, 107016, Москва, ул. Неглинная, 12, к. В).
E-mail: elena@lita.sh
https://orcid.org/0009-0006-9757-807X

Для цитирования: *Литаш-Сорокина Е.А.* Искусственный интеллект и дипфейки: вызовы и перспективы. *Государственная служба*. 2025. № 3. С. 37–50.

INFORMATION ABOUT THE AUTHOR:

Elena A. Litash-Sorokina, Head of the Center for Operations with Digital Rubles, Department of the National Payment System
The Central Bank of the Russian Federation (12, bldg. B, Neglinnaya St., Moscow, 107016, Russian Federation). E-mail elena@lita.sh
https://orcid.org/0009-0006-9757-807X

For citation: *Litash-Sorokina E.A.* Artificial Intelligence and deepfakes: challenges and prospects. *Gosudarstvennaya sluzhba*. 2025. No. 3. P. 37–50.