

Челябинский гуманитарий. 2024. № 3 (68). С. 24–35.

ISSN 1999-5407 (print).

Chelyabinskij Gumanitarij. 2024; 3 (68), 24–35.

ISSN 1999-5407 (print).

Научная статья

УДК 30;070

DOI 10.47475/1999-5407-2024-68-3-24-35

ДИПФЕЙК КАК ФЕНОМЕН СОВРЕМЕННОГО ИНФОРМАЦИОННОГО ПРОСТРАНСТВА

**Людмила Сергеевна Макарова¹, Юрий Викторович Баташев²,
Алексей Геннадьевич Солодовников³, Илья Валерьевич Померанцев⁴**

^{1,2,3,4} Национальный исследовательский Нижегородский государственный университет

им. Н. И. Лобачевского, Нижний Новгород, Россия

¹ iimakar@bk.ru, ORCID 0000-0001-8993-5475

² yury.batashev@gmail.com

³ solodovnikov.expert@outlook.com

⁴ ivpomerantsev@gmail.com

Аннотация. Для современного общества проблема циркуляции недостоверной (фейковой) информации приобретает критическое значение. Развитие и доступность технологий имеют и негативные последствия, включая возможность использования генеративного искусственного интеллекта (Gen AI) для создания мультимедийного фейкового контента (дипфейк), дезинформации и манипулирования общественным мнением. Современные технологии позволяют производить и распространять ложный мультимедийный контент, который порой неотличим от достоверных фактов и способен мгновенно распространяться в социальных сетях, а потом достаточно долго циркулировать в информационном пространстве. Ключевые факторы – это доступность и скорость распространения информации новостного и развлекательного характера, высокая степень доступности мобильных устройств и подключения к сети Интернет, доверие людей к мультимедийной информации и высказываниям публичных персон. Одним из триггеров массового появления мультимедийной фейковой информации после 2021–2022 гг. в сфере технологий является комодитизация (доступность) генеративного искусственного интеллекта (text2Image нейросети Stable Diffusion, DALL-E). Доступные технологии прикладного генеративного ИИ позволяют создавать на основе существующих данных различные виды мультимедийного контента – изображения, аудио и видео на любом современном устройстве (ПК, смартфон), в том числе подключая облачные вычислительные мощности и мультимодальные LLM, доступные в виде «чат-ботов» (ChatGPT, GigaChat). Одним из негативных аспектов создания сгенерированного и модифицированного контента становится его использование в мошеннических схемах. В исследовании приводятся описание методики создания дипфейков и примеры их функционирования в современном информационном пространстве, анализируются риски их использования. Идентификация фальшивого контента (дипфейков), созданного с помощью генеративного искусственного интеллекта, является сложной задачей, требующей разработки и внедрения эффективных алгоритмов и инструментов, способных обнаруживать и отличать подлинные материалы от недостоверных. В статье анализируются технологические аспекты данного процесса и актуальные разработки зарубежных и российских компаний, таких как Microsoft Video Authenticator, Intel FakeCatcher, ИС «Вепрь», ПО «Зефир», технологические решения ПАО «Сбербанк». Использование технологий искусственного интеллекта для создания дипфейков ставит перед обществом целый ряд важных этических вопросов, касающихся правды, законности и моральной приемлемости, права на частную жизнь, авторских прав, а также потенциальной угрозы для общественной безопасности. Это требует разработки и внедрения строгих этических норм и мер законодательного регулирования для предотвращения злоупотребления указанными технологиями, в том числе обязательная маркировка (watermark) сгенерированного Gen AI контента. В статье анализируется положительный опыт «Альянса в сфере искусственного интеллекта», который объединяет вузы и технологические компании России. В рамках деятельности организации разрабатываются рекомендации и этические документы с целью регулирования использования технологий ИИ, чтобы предотвратить злоупотребления и нарушения этических норм, в том числе и в сфере информации. Организация предлагает развивать и внедрять подходы и технологии, способные обнаруживать и отличать подлинные материалы от фальшивых, чтобы помочь гражданам и организациям критически оценивать контент, представленный в СМИ и социальных медиа.

Ключевые слова: дипфейк, верификация данных, фактчекинг, информационные технологии, искусственный интеллект, Альянс в сфере искусственного интеллекта, этика искусственного интеллекта.

Для цитирования: Макарова Л. С., Баташев Ю. В., Солодовников А. Г., Померанцев И. В. Дипфейк как феномен современного информационного пространства // Челябинский гуманитарий. 2024. № 3 (68). С. 24–35. doi: 10.47475/1999-5407-2024-68-3-24-35

Original article

DEEPFAKE AS A PHENOMENON OF THE MODERN INFORMATION SPACE

Lyudmila S. Makarova¹, Yuri V. Batashev², Alexey G. Solodovnikov³, Ilya V. Pomerantsev⁴

^{1, 2, 3, 4} Lobachevsky National Research Nizhny Novgorod State University, Nizhny Novgorod, Russia

¹ limakar@bk.ru, ORCID 0000-0001-8993-5475

² yury.batashev@gmail.com

³ solodovnikov.expert@outlook.com

⁴ ivpomerantsev@gmail.com

Abstract. The problem of the circulation of unreliable (fake) information in mass communications is of critical importance for modern society. The development and accessibility of technology has negative consequences, including the possibility of using generative artificial intelligence (Gen AI) to create fake multimedia content (deepfake), disinformation and manipulation of public opinion. Modern technologies allow you to create and distribute false multimedia content, which is sometimes indistinguishable from reliable facts and is able to instantly spread on social networks (go viral), and then circulate in the information space for a long time. The key factors are the availability and speed of dissemination of news and entertainment information, a high degree of penetration of mobile devices and Internet connectivity, people's trust in multimedia information and public figures. One of the triggers for the mass appearance of multimedia fake information after 2021–2022 in the field of technology is the commoditization (accessibility) of generative artificial intelligence (text2Image neural network Stable Diffusion, DALL-E). The available technologies of applied generative AI allow generating various types of multimedia content based on existing data – images, audio and video on any modern device (PC, smartphone), including connecting cloud computing power and multimodal LLMs available in the form of “chatbots” (ChatGPT, GigaChat, etc.). Generated and modified content is used in fraudulent schemes. The study describes the methodology for creating deepfakes and examples of their functioning in the modern information space, analyzes the risks of their use. Identifying fake content (deepfakes) created with the help of generative artificial intelligence is a difficult task that requires the development and implementation of effective algorithms and tools capable of detecting and distinguishing genuine materials from unreliable ones.

The article analyzes the technological aspects of this process and current developments of foreign and Russian companies, such as Microsoft Video Authenticator, Intel FakeCatcher, “Vepr” and “Zephyr” software, technological solutions of Sberbank PJSC. The use of artificial intelligence technologies to create deepfakes poses a number of important ethical questions to society regarding truth, legality and moral acceptability, the right to privacy, copyright, as well as a potential threat to public safety. This requires the development and implementation of strict ethical standards and legislative regulations to prevent the abuse of these technologies, including mandatory watermark of Gen AI generated content. The article analyzes the positive experience of the Alliance in the field of Artificial Intelligence, which unites universities and technology companies in Russia. As a part of the organization's activities, recommendations and ethical documents are being developed in order to regulate the use of AI technologies in order to prevent abuse and violations of ethical standards. The organization proposes to develop and implement approaches and technologies capable of detecting and distinguishing genuine materials from fake ones in order to help citizens and organizations critically evaluate information.

Key words: deepfake, data verification, fact checking, information technology, artificial intelligence, AI Alliance Russia, ethics of artificial intelligence.

For citation: Makarova L. S., Batashev Yu. V., Solodovnikov A. G., Pomerantsev I. V. (2024). Deepfake as a phenomenon of the modern information space. *Chelyabinskij Gumanitarij*, 3 (68), 24–35. doi: 10.47475/1999-5407-2024-68-3-24-35

Введение: постановка проблемы

Одним из наиболее актуальных и востребованных достижений в сфере технологий является искусственный интеллект (ИИ). Технологии прикладного ИИ позволяют анализировать и генерировать на основе существующих данных различные виды контента. Вместе с тем их развитие имеет и негативные последствия, включая возможность использования ИИ для создания фейкового контента и манипулирования общественным мнением. Технологии Gen AI (Generative Artificial Intelligence) и GAN (Generative Adversarial Network) также могут применяться в качестве инструмента в процессе создания и распространения дезинформации.

Проблема, связанная с функционированием дипфейков в современном информационном пространстве, включает следующие аспекты:

1. Технологический: идентификация фальшивого контента, созданного с помощью дипфейков, является сложной задачей, требующей разработки и внедрения эффективных алгоритмов и инструментов, способных обнаруживать и отличать подлинные материалы от недостоверных. В то же время следует отметить, что и сам процесс создания качественного дипфейка пока является достаточно трудоемким и требует профессиональных навыков.

2. Морально-этический аспект: использование технологий искусственного интеллекта для создания дипфейков ставит перед обществом целый ряд важных этических вопросов, касающихся моральной приемлемости таких действий. Данная тема вызывает всё больший интерес как в академической среде, так и у представителей государственной власти, ИТ-сообщества и общественности.

3. Правовые вопросы: как и любая технология, ИИ несет в себе определенные риски. В частности, это касается нарушения права на частную жизнь, защиты авторских прав, а также потенциальной угрозы для общественной безопасности. Доступность технологий позволяет использовать их в мошеннических схемах. Это требует разработки и внедрения законодательных регуляторов для предотвращения злоупотребления указанными технологиями. При этом на данный момент в России пока не существует целостного юридического подхода к степени «жесткости» данной регуляции, анализируется зарубежный опыт в данной сфере (США, Европейский союз, Китай), единства по этому вопросу в профессиональном ИТ-сообществе также пока нет.

3. Влияние на общественное мнение: дипфейки могут стать инструментом манипулирования общественным сознанием. Данные технологии позволяют создавать убедительные видео- и аудиоматериалы, которые могут быть ошибочно приняты за подлинные, что может привести к формированию неправильных представлений у большого количества людей, подрыву доверия к СМИ и политическим институтам, а также для распространения пропаганды и дезинформации, что может иметь серьезные последствия для общественной безопасности.

Материалы и методы исследования

Теоретические и практические аспекты применения актуальных технологий прикладного искусственного интеллекта в сфере журналистики и массовых коммуникаций являются предметом исследования зарубежных и отечественных авторов. Также существует большое количество работ, в которых анализируются вопросы, связанные с возможностями достижений в области ИТ, включая технологии генеративного ИИ, в системе проверки информации.

Следует отметить работы специалистов в области ИТ и массовых коммуникаций, анализирующих технологии создания и верификации дипфейков, а также последствия их применения в современном информационном пространстве: С. Bernaciak, D. Ross (Bernaciak, Ross 2022), S. Kh. Hamed, J. Ab Aziz Mohd, M. Ridzwan Yaakub (Hamed, Aziz Mohd, Ridzwan Yakoob 2023), K. Langmia (Langmia 2023), M. Lanhman (Lanhman 2021), D. Meredith (Meredith 2024), Ph. Pond (Pond 2020), C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, Ch. Bush (Rathgeb, Tolosana, Vera-Rodriguez, Bush 2022), N. Schick (Schick 2019).

Укажем исследования, в которых речь идет об описании непосредственно технологических аспектов процессов автоматизации процесса фактчекинга: В. С. Бережная (Бережная 2020), С. С. Распопова, С. И. Симакова (Распопова, Симакова 2022), P. Deepak, T. Chakraborty, Ch. Long, Santosh Kumar G (Deepak, Chakraborty, Long, G 2021), N. Hassan, Ch. Li, F. Arslan, M. Tremayne (Hassan, Li, Arslan, Tremayne 2017), N. Giansiracusa (Giansiracusa 2021), P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barron-Cedeno, P. Papotti, Sh. Shaar, G. Da San Martino (Nakov, Corney, Hasanain, Alam, Elsayed, Barron-Cedeno, Papotti, Shaar, Da San Martino 2021), S. Shaar, N. Georgiev, F. Alam, G. Da San Martino, A. Mohamed, P. Nakov (Shaar, Georgiev, Alam, Da San Martino, Mohamed, Nakov 2021). Анализ информационных систем, которые автоматизируют отдельные этапы процесса фактчекинга, а также перспективы создания эффективного инструмента проверки информации с помощью технологии нейросетей представлены в статье Л.С. Макаровой, Ю. В. Баташева (Макарова, Баташев 2023).

Данное исследование основано на использовании междисциплинарного подхода, который позволяет охарактеризовать различные аспекты дипфейка как феномена современного информационного пространства. В представленной статье анализируются характеристики технологических подходов к созданию дипфейков, а также возможности их верификации. Помимо этого, авторы, опираясь на метод сравнительного анализа, выявляют особенности существующих на данный момент специальных технологий, созданных с целью верифицировать визуальный и аудиоконтент, имеющий признаки дипфейка (Microsoft Video Authenticator, Intel FakeCatcher, ИС «Вепрь», ПО «Зефир», технологические решения ПАО «Сбербанк»). Также авторы статьи применяют в рамках исследования метод case study: в качестве эмпирического материала в работе представлены примеры дипфейков в России и за рубежом, включая использование технологий Gen AI и GAN в целях политической агитации в социальных медиа и как инструмента в руках мошенников. Выводы относительно возможностей этического ограничения использования технологий генеративного ИИ при создании дипфейков сформулированы на основе метода анализа, который был применен при изучении этических документов Альянса в сфере ИИ, в работе над созданием которых, равно как и в дискуссиях по данному вопросу, принимали участие непосредственно авторы статьи.

Результаты исследования

Создание убедительного дипфейка сегодня требует применения мощного графического процессора (GPU) игрового типа стоимостью около двух тысяч долларов США. Бесплатное программное обеспечение

с открытым исходным кодом для создания дипфейков легко найти, загрузить и установить. Однако комбинация серьезных навыков редактирования графики, видеомонтажа и работы со звуком, необходимых для получения правдоподобного результата, встречается нечасто. Более того, работа, необходимая для создания такого дипфейка, требует временных затрат в размере от нескольких дней до недель на обучение модели и устранение недостатков.

Двумя наиболее широко используемыми программными платформами с открытым исходным кодом для создания дипфейков на сегодняшний день являются DeepFaceLab и FaceSwap. Они общедоступны, поддерживаются крупными и организованными онлайн-сообществами с тысячами пользователей, многие из которых активно участвуют в разработке и повышении качества программного обеспечения и нейросетевых моделей. Это позволяет последовательно повышать удобство и совершенствовать инструментарий данных платформ, делая доступным создание правдоподобных поддельных видео и аудио даже для менее квалифицированных пользователей.

Из существующих подходов к созданию дипфейков можно выделить основанные на архитектуре энкодер-декодер и генеративно-сопоставительные сети (Generative Adversarial Network).

Процесс создания дипфейка с использованием схемы энкодер-декодер состоит из пяти этапов:

1. Сбор датасетов с видео цели (человека, лицо которого требуется подделать) и актера (человека, на лицо которого будет наложено лицо цели). Требуется как минимум несколько минут видеоматериала в качестве HD или 4K для актера и для цели. Видео должны демонстрировать схожие выражения лиц, движения глаз и повороты головы. Также актер и цель должны быть внешне практически идентичны.

2. Извлечение. На этом этапе каждое видео разбивается на кадры. В каждом кадре идентифицируется лицо (обычно с использованием модели DNN) и определяется около 30 ориентиров лица, которые служат опорными точками для модели, позволяющей определить местоположение черт лица.

3. Обучение. Каждый набор размеченных лиц затем вводится в обучающую сеть. Общая схема сети энкодер-декодер для обучения и преобразования показана на рисунке 1. Сгенерированные лица сравниваются с исходными лицами, вычисляется функция потерь, происходит обратное распространение и обновляются веса для сетей декодера и энкодера. Это необходимо для следующей серии лиц до тех пор, пока не будет достигнуто требуемое количество эпох (циклов обучения). «Пользователь сам решает, в какой момент прекратить обучение, визуально проверяя качество получаемых лиц, когда величина потерь больше не уменьшается. Бывают случаи, когда разрешение или качество входных изображений по разным причинам не позволяет значению потерь достичь желаемого показателя. Скорее всего, в этом случае никакое обучение или последующая обработка не приведут к созданию убедительного дипфейка» (Bernaciak C., Ross D., 2022: How Easy Is It to Make and Detect a Deepfake? Carnegie Mellon University, Software Engineering Institute's Insights (blog), accessed June 21, 2024, <https://insights.sei.cmu.edu/blog/how-easy-is-it-to-make-and-detect-a-deepfake/>).

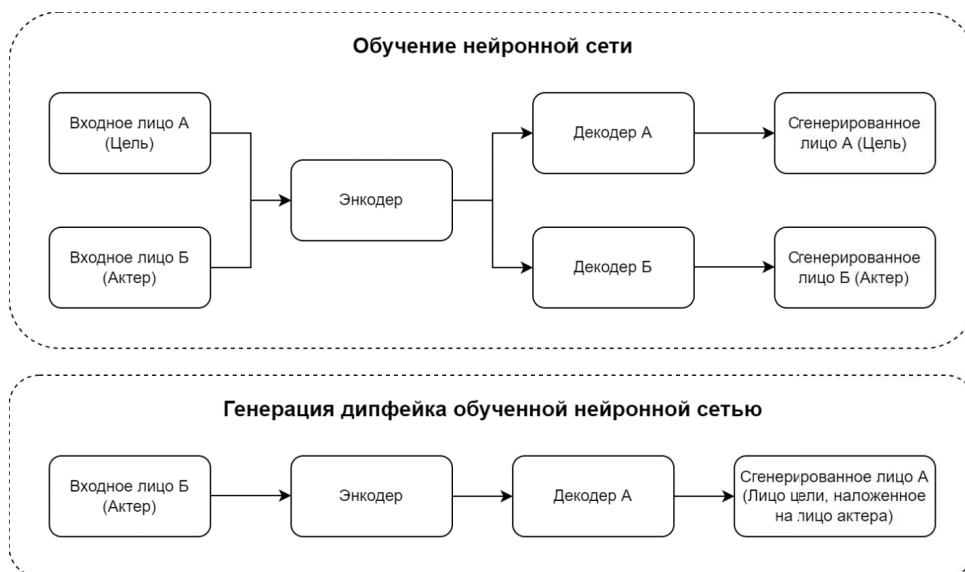


Рис. 1. Технология создания дипфейков с помощью схемы энкодер-декодер

4. Конвертация. На этапе конвертации создается непосредственно сам дипфейк. Если требуется заменить лицо Б лицом А, то используется схема, показанная в нижней части рисунка 1. Здесь размеченные лица Б подаются в энкодер. Напомним, что этот кодировщик обучен на размеченных представлениях для обоих

лиц А и Б. Когда выходные данные энкодера передаются в декодер для А, он попытается сгенерировать лицо А на основе лица Б. Этап конвертации представляет собой одностороннюю передачу набора входных данных по сети энкодер-декодер. Результатом процесса конвертации является набор кадров с замененными лицами, которые затем должны быть объединены другим программным обеспечением для преобразования в видео.

5. Пост-процессинг. Этот шаг требует значительного времени и навыков. Незначительные дефекты могут быть отредактированы, но значительные различия, скорее всего, не удастся устранить. Хотя постобработка может быть выполнена с использованием встроенных в программные платформы функций компоновки и маскировки, результаты, как правило, не слишком удовлетворительны.

Несмотря на то, что DeepFaceLabs предоставляет возможность поэтапной настройки цветокоррекции, положения маски, размера маски и ее растушевки для каждого кадра видео, их влияние на качество итогового результата довольно ограничено. Для достижения фотореалистичной постобработки требуются традиционные мультимедийные спецэффекты. Программная платформа для создания дипфейка на данном этапе используется только для экспорта полученного результата и набора параметров в традиционные видеоредакторы. ПО (программное обеспечение) DaVinci Resolve можно использовать для цветокоррекции и настройки цветности в соответствии с целевым видео. Затем ПО Mocha можно использовать для отслеживания планарного движения целевого видео, а также для выравнивания наложенного лица по ключевым кадрам. После этого результат из ПО Mocha можно импортировать в ПО Adobe After Effects для окончательной обработки.

Каждый инструмент создания дипфейков с открытым исходным кодом имеет большое количество настроек и гиперпараметров нейронной сети с некоторыми общими чертами между инструментами и некоторыми различиями, главным образом в отношении архитектуры нейронной сети. Благодаря широкому спектру доступных графических процессоров и возможности аренды облачных GPU, можно создавать качественный дипфейк с относительно небольшими финансовыми затратами на оборудование. Требования к оборудованию зависят от качества и разрешения целевого дипфейка. Наиболее важным аппаратным компонентом здесь является графический процессор. Он должен быть совместим с технологиями NVIDIA CUDA и фреймворком TensorFlow, для чего требуются графические процессоры компании NVIDIA.

«Современный уровень технологий создания дипфейков предполагает длительный процесс сбора или записи исходного видеоматериала, обучение нейронных сетей, метод проб и ошибок для поиска наилучших параметров и тщательную постобработку видео. Поэтому создание дипфейка – это своего рода работа на стыке искусства и науки». (Bernaciak C., Ross D., 2022: How Easy Is It to Make and Detect a Deepfake? Carnegie Mellon University, Software Engineering Institute's Insights (blog), accessed June 21, 2024, <https://insights.sei.cmu.edu/blog/how-easy-is-it-to-make-and-detect-a-deepfake/>).

Генеративно-сопоставительные сети (GAN) являются второй популярной на сегодняшний день технологией создания дипфейков. Цель GAN – создать что-то новое на основе предыдущих данных. Например, она может воссоздать человеческое лицо после изучения сотен фотографий. Или же она может создать картину, напоминающую стиль определенного художника, используя его работы в качестве опорного материала. В подходе GAN две нейронные сети – генератор и дискриминатор – настраиваются на прямую конкуренцию друг с другом. Генератор выдает новое изображение в качестве выходных данных на основе знаний, на которых была обучена нейронная сеть. Дискриминатор определяет, является ли изображение реальным или поддельным. Оба компонента находятся в постоянном взаимодействии. Генератор учится создавать изображения, которые «обманут» дискриминатор и заставят его классифицировать созданное изображение как реальное. Дискриминатор, с другой стороны, «учится» не поддаваться обману.

Чем лучше дискриминатор, тем сложнее генератору будет создавать реалистичные изображения и в конечном счете тем лучше он будет выполнять свою работу. В каком-то смысле здесь подходит в качестве аналогии процесс передачи знания от учителя к ученику: чем лучше преподаватель (дискриминатор) проводит обучение, тем больше это помогает ученику (генератору) достигать хороших результатов. Ученик представляет свою работу, а учитель отмечает его ошибки: только так ученик сможет осознать, что он сделал неправильно, и исправить недочеты.

Распространение дипфейков в сфере массовой коммуникации на данный момент постоянно увеличивается. В отдельных случаях это вызывает вопросы не только в контексте правил ведения общественно-политической дискуссии, но и в сфере морали. Например, 6 января 2024 г. во время предвыборной кампании в Индонезии заместитель председателя партии «Голкар» Эрвин Акса распространил в соцсетях дипфейк-видео, в котором ныне покойный индонезийский президент Хаджи Сухарто призвал голосовать за кандидата от партии «Голкар». Данный дипфейк стал вирусным, набрав за пять дней 4,2 миллиона просмотров и 1200 комментариев, и вызвал дебаты об этических и юридических последствиях использования таких технологий. (<https://www.youtube.com/watch?v=DOu8VlxNyFA>). Проблема функционирования дипфейков в системе политической коммуникации и ее правовые и морально-

этические аспекты обсуждались в рамках совместных онлайн-семинаров Университета Лобачевского и Джакартского университета развития «Ветеран» «Медиасистемы России и Индонезии: движение навстречу друг другу», которые проводились 23 апреля и 29 мая 2024 г. и были посвящены обсуждению актуальных проблем медиа в обеих странах (<https://int.unn.ru/news/seminar-mediasistemy-rossii-i-indonezii-dvizhenie-navstrelu-drug-drugu-proshyol-v-nngu/>).

В России 25 июня 2024 г. Пленум Верховного Суда РФ дал разъяснения относительно использования технологий дипфейков в агитационных материалах в ходе избирательных кампаний, указав, что это «является нарушением требований к агитационным материалам, за которое предусмотрена административная ответственность» (Постановление Пленума Верховного Суда РФ №17 от 25.06.2024 «Об отдельных вопросах, возникающих у судов при рассмотрении дел об административных правонарушениях, посягающих на установленный порядок информационного обеспечения выборов и референдумов»: «Привлечению к административной ответственности по части 1 статьи 5.12 КоАП РФ подлежат также заказчики и лица, выполнившие работы, оказавшие услуги по созданию (подготовке) агитационных материалов всех видов, в частности, со следующими нарушениями: ...с использованием вводящих в заблуждение и выдаваемых за достоверные недостоверных изображений, аудио- и аудиовизуальной информации, в том числе созданных с помощью компьютерных технологий (пункт 11 статьи 56 Закона об основных гарантиях избирательных прав)»; <https://www.vsr.ru/documents/own/33741/>).

В качестве примера практически повсеместного применения технологии дипфейка (в том числе как развлечения) можно привести сюжет городского портала «Открытый Нижний». В декабре 2023 г. журналисты применили технологию клонирования голоса в сочетании с технологией синхронизации lip sync. Они сделали так, чтобы губернатор Нижегородской области Глеб Никитин, глава Нижнего Новгорода Юрий Шалабаев и другие политики и общественные деятели Нижегородской области «исполнили» песню «В лесу родилась елочка». Сотрудники редакции клонировали голоса представителей власти и с помощью lip sync синхронизировали движение их губ на фотографиях, сделав так, чтобы они двигались в такт вокалу (Telegram-канал городского портала «Открытый Нижний». Режим доступа: <https://t.me/opennov/16069>).

Хотя подавляющее большинство создаваемых дипфейков пока имеют развлекательный характер, вызывает тревогу растущее число использования данной технологии для реализации сценариев мошенничества. Например, по данным издания South China Morning Post, в начале 2024 г. британская инженерная компания Acqr в Гонконге стала жертвой мошенничества с использованием технологии дипфейка, потеряв 25,6 млн долларов. Сотрудников местного филиала обманули с помощью цифрового двойника главного финансового директора, который в ходе видеоконференции приказал провести денежные переводы. В отличие от предыдущих случаев мошенничества, где жертв обманывали в ходе видеозвонков в формате «один-на-один», злоумышленники использовали технологию дипфейка для создания групповых видеоконференций, где внешность и голоса мошенников были неотличимы от реальных руководителей. Такой подход позволил обмануть сотрудника финансового отдела. Он получил текстовое сообщение якобы от главного финансового директора компании из Великобритании о необходимости секретной транзакции и, несмотря на первоначальные сомнения, принял участие в групповой видеоконференции, где «присутствовали» директор, другие работники компании и сторонние лица. Их голоса и видео были почти неотличимы от настоящих. Сотрудник следовал инструкциям собеседников и осуществил 15 переводов на общую сумму 200 миллионов гонконгских долларов на 5 банковских счетов в Гонконге. (<https://www.scmp.com/news/hong-kong/law-and-crime/article/3250851>).

Вопрос о противодействии практике мошеннических действий в рамках использования технологии дипфейка остро стоит и в России. «Развитие компьютерных технологий привело к расширению возможностей по созданию видео- и аудиоматериалов на основе образцов изображений и голоса человека, искусственно воссоздающих несуществующие события. Современные программно-аппаратные комплексы, а также использование нейросетей и искусственного интеллекта (технологии “дипфейк”, “цифровые маски” и так далее) позволяют создавать подделки, отличить которые от реальности неспециалисту практически невозможно», – говорится в пояснительной записке к законопроекту о внесении поправок в статьи УК РФ «Клевета», «Кража», «Мошенничество», «Вымогательство», «Причинение имущественного ущерба путем обмана или злоупотребления доверием» относительно введения уголовного наказания за использование изображения или голоса, а также биометрических данных граждан без их согласия, которые должен в ближайшее время представить в ГД РФ глава Комитета по труду, социальной политике и делам ветеранов Ярослав Нилов (Башлыкова Н., Крылова Е. Голосовые связи: за дипфейки хотят ввести уголовную ответственность. Известия. 2024. Май 24., <https://iz.ru/1702846/natalia-bashlykova-elizaveta-krylova/golosovye-sviasi-za-dipfeiki-khotiat-vvesti-ugolovnuu-otvetstvennost>). На данный момент законопроект проходит стадию согласования в Правительстве РФ и в Верховном Суде, после чего будет внесен в Государственную Думу.

Определение дипфейкового контента основывается на формате медиаматериалов (видео, аудио,

изображение) и временном контексте, в котором субъект дипфейка (объект подделки) представлен в медиапространстве. В режиме прямой трансляции (NRT – near real time) зрители и слушатели имеют возможность задавать вопросы, требующие детальной информации о профессиональной деятельности или личной жизни субъекта. Обычно такие данные быстро проверяются онлайн. Второй фактор, который следует учитывать, это технические сложности одновременной подделки видео и аудио. Для этого необходимо создать дополнительный канал и внести резкие изменения в условия его функционирования. Например, можно попросить говорящего более громко произносить слова, четче или издавать резкие звуки. Аналогично, для видео можно изменить условия освещения, фокусное расстояние или наличие окружающих предметов. Важно отметить, что технологии обычно хорошо работают при фронтальной проекции, но испытывают трудности при боковом ракурсе. Эксперты, анализирующие контент постфактум, исследуют наличие аномалий (звуки, уровни, яркость, пульсации и т. д.), классифицируют артефакты (предметы, щелчки, посторонние звуки и т. д.) и оценивают степень «идеальности» происходящего. Человек воспринимает «идеальность» через органы чувств, включая слух и зрение, как созданное искусственно. Данное явление, получившее название «эффект зловещей долины», обсуждается в том числе и в рамках научно-популярной проблематики в медиа (Палихова А. Что такое эффект зловещей долины и почему нас пугают роботы // РБК. 2024. Февраль 22, <https://trends.rbc.ru/trends/futurology/60f179059a794715da3bd9ba>).

Следует указать следующие методы обнаружения дипфейкового контента:

1. Анализ шумов и артефактов — это метод обнаружения дипфейков, основанный на поиске аномальных шумов и артефактов в изображении. Подобные аномалии могут быть вызваны ошибками в процессе генерации дипфейка или намеренно добавлены для создания эффекта реалистичности. Для обнаружения этих аномалий используются специализированные алгоритмы, которые анализируют структуру изображения и сравнивают ее с ожидаемыми характеристиками. Если обнаруживаются отклонения от нормы, то это может свидетельствовать о вероятности наличия дипфейка.

2. Анализ текстур и деталей: метод основан на сравнении изображений, которые вызывают подозрение на генерацию дипфейка, с реальными изображениями людей.

3. Сравнение с базой данных: подход основан на том предположении, что большинство реальных изображений уже наличествуют в базах данных сети Интернет. Если проверяемое изображение совпадает по определенным параметрам, то это чаще всего является свидетельством его подлинности. Отсутствие такого совпадения, скорее всего, указывает на дипфейк. Отметим базы различных изображений (датасеты), применяемые для целей машинного обучения, например, CIFAR-100, ImageNet, LSUN, MS COCO, Places, Google Open Images (30 миллионов изображений), (<https://storage.googleapis.com/openimages/web/index.html>). Аналогичный подход применяется и в процессе выявления дипфейковых аудиоматериалов.

Также следует отметить еще один эффективный метод обнаружения, например, дипфейковых видео: анализ синхронизации между аудио- и видеодорожками в ролике. Несовпадение между ними в процессе просмотра видео является одним из важных маркеров того, что речь идет о генерации контента. Помимо данного аспекта, следует отметить, что в процессе верификации дипфейкового видео важно проанализировать окружающий фон. Для этого существуют специальные инструменты, например, представленные на сайте <https://deerware.ai/>. Проверка аудиодипфейков связана с анализом несоответствий или аномалий в фоновых звуках, что может свидетельствовать в пользу того, что аудиозапись является результатом генерации.

Следует отметить существующие на данный момент специальные технологии, призванные верифицировать визуальный и аудиоконтент, имеющий признаки дипфейка. Например, **Intel FakeCatcher** — это инструмент на базе технологий OpenVino, OpenCV, Deep Learning Boost и других, разработанный Intel для обнаружения дипфейков в реальном времени. По оценке Intel, инструмент способен определять фальшивые видео с точностью до 96 %. FakeCatcher работает путем анализа пульсаций света, который поглощается и отражается кожей человека в зависимости от фазы кардиоцикла. Измеряя количество света, которое поглощается и отражается 32 точками на лице человека, создается карта фотоплетизмограммы (PPG). Пространственно-временная фотоплетизмограмма позволяет оценить естественную природу лица на видео. Подобные подходы применяют в оптических пульсоксиметрах (<https://www.intel.com/content/www/us/en/newsroom/news/intel-introduces-real-time-deepfake-detector.html>). На сегодня данная технология не размещена в открытом доступе.

Компания Microsoft еще в 2020 г. заявила о начале разработки системы распознавания дипфейков. Инструмент под названием **Microsoft Video Authenticator** использует специальный алгоритм для определения подлинности какого-либо изображения или видеоролика. На сайте компании представлена развернутая статья (за авторством Тома Берта, корпоративного вице-президента по безопасности и защите клиентов, и Эрика Хорвица, главного научного сотрудника Microsoft), в которой описывается подход к созданию технологии верификации дипфейковых изображений: «Технология была разработана командой Microsoft Research в сотрудничестве с Microsoft Responsible AI и Комитетом Microsoft по ИИ, этике и

последствиям в разработке и исследованиях (Microsoft AI, Ethics and Effects in Engineering and Research, AETHER)». В качестве партнеров в сфере медиа в статье были указаны BBC, CBC/Radio-Canada и New York Times, а также USA Today. (<https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/>).

Следует отметить, что о результатах создания данной технологии в настоящее время в открытом доступе нет информации. Однако в апреле 2024 г. на сайте Microsoft и впоследствии в СМИ появилась информация о нейросети VASA-1, которая может создавать видео на основе одного изображения. Для использования нужно предоставить картинку и аудиодорожку, и алгоритм сгенерирует говорящего человека с естественной мимикой и широким спектром эмоций (<https://www.microsoft.com/en-us/research/project/vasa-1/>; Аверьянова В. Цифровой близнец: Microsoft представила ИИ для создания реалистичных видео. Нейросеть делает ролики по фото и аудио // Известия. 2024. Апрель 30, <https://iz.ru/1684945/>).

Российский опыт создания технологии верификации информации представлен в рамках деятельности АО «Крибрум» (в декабре 2022 г. компания получила грант Российского фонда развития информационных технологий (РФРИТ) на развитие проекта «Разработка системы анализа и визуализации разрозненных данных, включая данные социальных сетей «Крибрум. OSINT»). Согласно информации, размещенной на сайте компании, разработка указанного ПО на данный момент успешно завершена (<https://kribrum.ru/technology>).

Также следует отметить проекты под эгидой Федеральной службы по надзору в сфере связи, информационных технологий и массовой коммуникации (Роскомнадзор) и АНО «Диалог Регионы». Например, по заданию Роскомнадзора разработана ИС «Вебрь»: «Разработка информационной системы “Вебрь” ведется с 2022 года. “Вебрь” вместе с автоматизированной системой “Окулус”, которая выявляет нарушения законодательства РФ в изображениях и видеоматериалах, входит в единую систему мониторинга информационного пространства» (<https://tass.ru/obschestvo/17091419>). Разработчиком данного ПО является ИТ-компания из Санкт-Петербурга «Необит» (<https://www.ixbt.com/news/2023/02/20/v-rossii-gotovjat-k-zapusku-sistemu-vepr-dlja-obezvrezhivaniya-infobomb-v-internete.html>), техническое задание предполагает, что «Вебрь» будет бороться с «распространением общественно значимой информации под видом достоверных сообщений, которая создает угрозу причинения вреда жизни и (или) здоровью граждан, имуществу, угрозу массового нарушения общественного порядка и (или) общественной безопасности». На сайте компании нет информации о данном проекте (<https://neobit.ru/products>).

АНО «Диалог регионы» в 2023 г. запустил информационную систему мониторинга аудиовизуальных материалов на основе распознавания речи в режиме реального времени. Соответствующая разработка представлена в ходе сессии ПМЭФ-2023 «Устойчивый цифровой диалог как драйвер многополярного мира». ПО «Зефир» работает как информационная система мониторинга аудиовизуальных материалов на основе транскрипции в режиме реального времени. Она позволяет моментально выявлять дипфейки благодаря алгоритмической оценке и анализу с помощью искусственного интеллекта. «Зефир» позволяет верифицировать материалы категории Deepfake с точностью результатов 77,3%» (<https://dialog.info/ustojchivuj-cifrovuj-dialog-rossii-i-miru/>).

ПАО «Сбербанк» в 2022 г. получил два патента от Федеральной службы по интеллектуальной собственности, которые относятся к технологиям, разработанным в ходе исследовательского проекта по идентификации дипфейков. Основной целью данного проекта является улучшение точности и эффективности обнаружения синтетических изменений в изображениях лиц людей, представленных в видеоматериалах. Данная информация и описание изобретений к патентам размещены на сайте Федеральной службы по интеллектуальной собственности (Роспатент).

В описании указано, что «технологическая основа включает в себя наборы ансамблей нейронных сетей, принадлежащих к классу EfficientNet (патент № 2768797), а также метод усиления и анализа с помощью нейронных сетей микроизменений в цветовых характеристиках объектов на отдельных кадрах (патент № 2774624). При объединении этих методов в единую систему они обеспечивают высокую точность определения синтетически измененных изображений лиц на видео. Алгоритм усиления Эйлера, применяемый при данном подходе, основан на манипуляциях с R (красной) компонентой RGB-изображения, предполагая, что видео с живым человеком содержит паттерны пульсаций крови в коже, коррелирующие с кардиоритмом» (Сбер запатентовал технологии по распознаванию дипфейков // Федеральная служба по интеллектуальной собственности (Роспатент). 2022. Август. 17, <https://rospatent.gov.ru/ru/news/sber-dipfejk-17082022>, ссылки на описание патентов https://www1.fips.ru/registers-doc-view/fips_servlet?DB=RUPAT&DocNumber=2768797&TypeFile=html, https://www1.fips.ru/registers-doc-view/fips_servlet?DB=RUPAT&DocNumber=2774624&TypeFile=html).

Следует отметить, что создатели технологии VASA-1, о которой шла речь выше, пообещали, что она не будет представлена в открытом доступе, поскольку существует серьезный риск ее использования в преступных целях (<https://www.microsoft.com/en-us/research/project/vasa-1/>). Действительно, дипфейк-

технологии, позволяющие создавать фальшивые видеозаписи и изображения, стали значительным вызовом для современной журналистики и общества в целом. Развитие технологий предполагает объединение усилий государства и общества в формировании инструментов их регулирования. В рамках ПМЭФ-24 7 июня 2024 г. состоялась дискуссия «Международный и национальный опыт регулирования сферы искусственного интеллекта. Наилучшие практики». Председатель комитета ГД по информационной политике, информационным технологиям и связи Александр Хинштейн, который принял участие в обсуждении, отметил, что правовое регулирование ИИ в России должно быть «мягким»: «Перед глазами обширный зарубежный опыт, и уже можно видеть прямую закономерность, что в странах, где эта сфера жестко зарегулирована, развитие ИИ идет медленно. Россия не должна повторять эту ошибку — мягкое регулирование сможет помочь развивать генеративные технологии. Уже есть реальные примеры самоорганизации разработчиков, вроде “Кодекса этики в сфере ИИ” и работы “Комиссии по этике в сфере ИИ”» (<https://forumspb.com/programme/business-programme/131535/>).

«Альянс в сфере искусственного интеллекта (ИИ)», результаты работы которого упоминает в своем выступлении А. Е. Хинштейн, объединяет ведущие технологические компании и вузы России (ННГУ им. Н. И. Лобачевского является членом организации) и при разработке рекомендаций по использованию и развитию компетенций ИИ в образовании, науке и инновациях всегда опирается на международный опыт (<https://a-ai.ru>).

На данный момент членам Альянса рекомендуется создавать независимые органы, ответственные за мониторинг и регулирование использования ИИ и дипфейков в частности, чтобы предотвратить злоупотребления и нарушения этических норм. Несомненно, что опыт саморегулирования в рамках профессионального сообщества, заинтересованность прежде всего разработчиков в сфере ИТ в вопросах этического ограничения развития технологий прикладного ИИ (об этом свидетельствует активное участие представителей ИТ-индустрии в обсуждении моральных критериев и обязанностей в контексте развития ИТ) может стать основой регулирования процесса развития информационных технологий, в том числе и в области права.

Альянс предлагает развивать и внедрять технологии, способные обнаруживать и отличать подлинные материалы от фальшивых, в том числе созданных с помощью технологии дипфейка, чтобы помочь гражданам и организациям критически оценивать информацию и не поддаваться на манипуляции. Для этих целей под эгидой организации разработаны следующие документы и рекомендации:

- Кодекс этики в сфере искусственного интеллекта (26.10.2021 г.);
- Этические рекомендации по применению рекомендательных технологий и алгоритмов, основанных на искусственном интеллекте, в цифровых сервисах (19.09.2023 г.);
- Рекомендации Комиссии по реализации Кодекса этики в сфере ИИ по теме создания и использования цифровых имитаций (10.04.2024 г.).

Основная задача деятельности Альянса в сфере ИИ – разработать рекомендации, которые будут способствовать развитию инструментов на базе ИИ и дадут понятные правила этичного использования таких инструментов пользователям. Многие консервативно настроенные эксперты выступают за жесткий контроль использования систем на базе ИИ. Те, кто уже непосредственно разрабатывают подобные технологии, наоборот, видят в них потенциал технологического и социального развития общества.

В то же время необходимо учитывать риск дискриминации и предвзятости, которые могут быть встроены в алгоритмы ИИ. Это может привести к несправедливому обращению с отдельными группами населения, что нарушает основные принципы равенства и справедливости. В связи с этим разработка этических рекомендаций основывается на строгих критериях прозрачности, объяснимости и ответственности за действия систем на базе ИИ.

Комиссия по этике призвана ответить на множество вопросов, с которыми будут сталкиваться разработчики и простые пользователи. В документе «Кодекс этики в сфере искусственного интеллекта» (разработан под эгидой Альянса в сфере искусственного интеллекта, 2023) подчеркивается: «Актеры ИИ не должны допускать передачи полномочий ответственного нравственного выбора СИИ (Системы искусственного интеллекта), делегировать ответственность за последствия принятия решений СИИ – за все последствия работы СИИ всегда должен отвечать человек (физическое или юридическое лицо, признаваемое субъектом ответственности согласно действующему законодательству РФ)» (<https://ethics.a-ai.ru>). Таким образом, ответственность за верификацию достоверности сгенерированного контента лежит на человеке, который его публикует и распространяет. Это означает, что именно он должен выполнить соответствующие обязанности по проверке контента перед его публикацией, а также нести ответственность за последствия.

Альянс в сфере ИИ проводит кампанию по информированию и просвещению общественности относительно рисков и опасностей, связанных с дипфейками, а также важности применения критического мышления в системе проверки информации. С данной целью в настоящее время разрабатывается проект

«Белая книга. 100 ответов на вопросы об этике искусственного интеллекта», где будут даны разъяснения относительно самых популярных вопросов по использованию ИИ, в том числе и касательно дипфейков, а также обязанностей по верификации достоверности сгенерированного контента. И. В. Померанцев, член Комиссии по этике Альянса в сфере ИИ, и Л. С. Макарова, член рабочей группы по этике ИИ в образовании Комиссии по этике ИИ, участвуют в проекте в качестве авторов.

Заключение

Несомненно, что стремительное развитие технологий и связанные с этим процессом риски вызывают опасения. На современном этапе технологии создания дипфейков развиваются быстрее, чем возможности их верификации. Результаты анализа кейсов, представленных в статье, позволяют сделать выводы относительно возможностей использования дипфейка практически во всех сферах жизни, включая негативные. Технологии генеративного ИИ в процессе постоянного совершенствования становятся эффективным инструментом создания контента, связанного с различными аспектами окружающей действительности, включая и негативные. В то же время представленный в статье анализ существующих на данный момент инструментов верификации видео- и аудиоконтента, имеющего признаки использования технологий Gen AI и GAN (Microsoft Video Authenticator, Intel FakeCatcher, ИС «Вепрь», ПО «Зефир», технологические решения ПАО «Сбербанк»), свидетельствует о том, что они практически недоступны для массовой аудитории, находятся либо в закрытом доступе, либо предполагают функционирование только в формате платной услуги.

В сложившейся ситуации главным инструментом в противодействии негативным последствиям распространения технологии дипфейка становится ужесточение законодательных норм (характеристика законодательных инициатив приводится в статье). Однако на данный момент и многие представители власти (мнение депутата ГД А. Е. Хинштейна, приведенное в исследовании), и разработчики в полной мере к этому не готовы.

Как уже отмечалось выше, эксперты Комиссии по этике ИИ придерживаются того мнения, что ответственность за верификацию сгенерированного при помощи технологий ИИ контента лежит на человеке, который ее публикует и распространяет, что предполагает и выполнение соответствующих обязанностей по его проверке. Понимание этого факта повышает ответственность за использование и распространение сгенерированного с помощью технологий ИИ контента. Данный вывод в наибольшей степени относится к вопросу создания и распространения дипфейков и других видов ложной информации и их последствий для общественной безопасности, что предполагает объединение усилий разработчиков в сфере технологий ИИ, массовой коммуникации, этики, права в сфере искусственного интеллекта в системе противодействия негативным последствиям этого явления.

Список источников

Бережная В. С. Вопросы стандартизации фактчекинга в журналистике данных. Теоретический аспект // Наука телевидения. 2020, № 16.2. С. 191–209.

Макарова Л. С., Баташев Ю. В. Перспективы использования технологий прикладного искусственного интеллекта в системе верификации информации СМИ и социальных медиа // Знак: проблемное поле медиаобразования. 2023. № 2 (48). С. 118–127.

Распопова С. С., Симакова С. И. Фактчекинг как новый формат саморегулирования сетевой коммуникации // Знак: проблемное поле медиаобразования. 2022. № 1 (43). С. 150–157.

Bernaciak C., Ross D. How Easy Is It to Make and Detect a Deepfake? Carnegie Mellon University, Software Engineering Institute's Insights (blog), accessed June 21, 2024, <https://insights.sei.cmu.edu/blog/how-easy-is-it-to-make-and-detect-a-deepfake/>.

Botnevik B., Sakariassen E., Setty V. BRENDA: Browser Extension for Fake News Detection. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China. ACM, New York, NY, USA, P. 1–4. <https://dl.acm.org/doi/10.1145/3397271.3401396>

Deepak P., Chakraborty T., Long Ch., Santosh Kumar G. DATA Science for Fake News. Surveys and Perspectives. Springer, 2021, 308 p.

Giansiracusa N. How Algorithms create and prevent fake news. Exploring the impacts of social media, deepfakes, GPT-3, and more. Apress, 2021.239 p.

Hamed S.Kh., Ab Aziz Mohd J., M. Ridwan Yaakub. A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion. Helyon, vol.9, issue 10, 2023, <https://doi.org/10.1016/j.heliyon.2023.e20382>.

Hasimi L., Poniszewska-Marañda A. Browser Extension for Detection of Fake News and Disinformation. In: Papadaki, M., Rupino da Cunha, P., Themistocleous, M., Christodoulou, K. (eds) Information Systems. EMCIS 2022. Lecture Notes in Business Information Processing, vol. 464. Springer, 2022, <https://doi.org/10.1007/978-3->

031-30694-5_16

Hassan N., Li Ch., Arslan F., Tremayne M. Toward Automated Fact-Checking: Detecting Check Worthy Factual Claims by ClaimBuster // KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13–17, 2017. New York: Association for Computing Machinery, 2017. P. 1803–1812.

Langmia K., *Black Communication in the Age of Disinformation. DeepFakes and Synthetic Media*. Palgrave Macmillan, 2023. 503p.

Lanhman M., *Generating a New Reality: From Autoencoders and Adversal Networks to Deepfakes*. Calgary, Canada, 2021. 327 p.

Meredith D., *The OSINT Handbook. A practical guide to gathering and analyzing online information*. Pact Publishing, 2024. 332 p.

Miranda S., Nogueira D., Mendes A., Vlachos A. Automated Fact Checking in the News Room. In Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA. ACM, New York, NY, USA, P. 1–6. <https://dl.acm.org/doi/10.1145/3308558.3314135>

Nakov P., Corney D., Hasanain M., Alam F., Elsayed T., Barron-Cedeno A., Papotti P., Shaar Sh., Da San Martino G. Automated Fact-Checking for Assisting Human Fact. Doxa: Qatar Computing Research Institute, 2021. 250 p.

Pond Ph. *Complexity, Digital Media and Post Truth Politics. A Theory of Interactive Systems*. Palgrave Macmillan, 2020. 247 p.

Rathgeb C., Tolosana R., Vera-Rodriguez R., Bush Ch. *Handbook of Digital Fact Manipulation and Detection. From DeepFakes to Morphing Attacks*. Springer, 2022. 358 p.

Shaar S., Georgiev N., Alam F., Da San Martino G., Mohamed A., Nakov P. Assisting the Human Fact-Checkers: Detecting All Previously Fact-Checked Claims in a Document. Doxa: Qatar Computing Research Institute, 2021. 308 p.

Shick N. *Deepfakes. The Coming Infocalypse*. New York- Boston, 2020. 436 p.

References

Berezhnaya, V. S. (2020). Voprosi standartizatsii faktchekinga v journalistike dannyh. Teoreticheskiy aspekt [Standardized Fact Checking in Data Journalism. A Theoretical view]. *Nauka televideniya*, 16.2, 191–209. (In Russ.).

Makarova, L. S. & Batashev, Yu. V. (2023). Perspektivy ispolzovaniya tekhnologii prikladnogo iskusstvennogo intellekta v sisteme verifikatsii informatsii SMI i socialnykh media [Prospects for the use of Applied Artificial Intelligence technologies in the verification system of media and Social Media information]. *Znak: problemnoe pole mediaobrazovaniya*, 2 (4), 118–127. (In Russ.).

Raspopova, S. S. & Simakova, S. I. (2022). Faktcheking kak noviy format samoregulirovaniya setvoy kommunikatsii. [Fact-checking as a new format of self-regulation network communication]. *Znak: problemnoe pole mediaobrazovaniya*, 1 (43), 150–157. (In Russ.).

Bernaciak, C. & Ross, D. (2022). How Easy Is It to Make and Detect a Deepfake? *Carnegie Mellon University, Software Engineering Institute's Insights (blog)*, available at: <https://insights.sei.cmu.edu/blog/how-easy-is-it-to-make-and-detect-a-deepfake/>. (accessed: 21.06.2024).

Botnevik, B., Sakariassen, E. & Setty, V. (2020). BRENDA: Browser Extension for Fake News Detection. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China*. ACM, New York, NY, USA, P. 1–4, available at: <https://dl.acm.org/doi/10.1145/3397271.3401396> (accessed: 24.06.2024).

Deepak, P., Chakraborty, T., Long, Ch. & Santosh Kumar G. (2021). *DATA Science for Fake News. Surveys and Perspectives*. Springer, 308 p.

Giansiracusa, N. (2021). *How Algorithms create and prevent fake news. Exploring the impacts of social media, deepfakes, GPT-3, and more*. Apress, 239 p.

Hamed S.Kh., Ab Aziz Mohd J. & M. Ridzwan Yaakub. *A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion*. Helyon, vol.9, issue 10, 2023, <https://doi.org/10.1016/j.heliyon.2023.e20382> (accessed: 25.06.2024).

Hasimi, L. & Poniszewska-Marañda, A. (2022). *Browser Extension for Detection of Fake News and Disinformation*. Papadaki, M., Rupino da Cunha, P., Themistocleous, M., Christodoulou, K. (eds) *Information Systems. EMCIS 2022. Lecture Notes in Business Information Processing*, vol. 464. Springer, available at: https://doi.org/10.1007/978-3-031-30694-5_16 (accessed: 24.06.2024).

Hassan, N., Li, Ch., Arslan, F. & Tremayne, M. (2017). *Toward Automated Fact-Checking: Detecting CheckWorthy Factual Claims by ClaimBuster*. KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13–17, 2017. New York: Association for Computing Machinery. 1803–1812.

Langmia K. (2023). *Black Communication in the Age of Disinformation. DeepFakes and Synthetic Media*. Palgrave Macmillan. 503p.

Lanhman, M. (2021). *Generating a New Reality: From Autoencoders and Adversal Networks to Deepfakes*. Calgary, Canada, 327 p.

Meredith, D. (2024). *The OSINT Handbook. A practical guide to gathering and analyzing online information*. Pact Publishing, .332 p.

Miranda, S., Nogueira, D., Mendes, A., Vlachos, A. (2019). *Automated Fact Checking in the News Room. Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA*. ACM, New York, NY, USA, available at: <https://dl.acm.org/doi/10.1145/3308558.3314135> (accessed: 25.06.2024).

Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barron-Cedeno, A., Papotti, P., Shaar, Sh., Da San Martino, G. (2021). *Automated Fact-Checking for Assisting Human Fact*. Doxa: Qatar Computing Research Institute, 250 p.

Deepak, P., Chakraborty, T., Long, Ch., Santosh, Kumar G. (2021). *DATA Science for Fake News. Surveys and Perspectives*. Springer, 308 p.

Pond, Ph. (2020). *Complexity, Digital Media and Post Truth Politics. A Theory of Interactive Systems*. Palgrave Macmillan, 247 p.

Rathgeb C., Tolosana R., Vera-Rodriguez R., Bush Ch. (2022). *Handbook of Digital Fact Manipulation and Detection. From DeepFakes to Morphing Attacks*. Springer, 358 p.

Shaar, S., Georgiev, N., Alam, F., Da San Martino, G., Mohamed, A., Nakov, P. (2021). *Assisting the Human Fact-Checkers: Detecting All Previously Fact-Checked Claims in a Document*. Доха: Qatar Computing Research Institute, 308 p.

Shick, N (2020). *Deepfakes. The Coming Infocalypse*. New York- Boston, 436 p.

Информация об авторах

Л. С. Макарова – кандидат филологических наук, доцент, доцент кафедры журналистики, заместитель по методической работе директора Института филологии и журналистики.

Ю. В. Баташев – системный аналитик компании «БИ-ТЕЛЕКОМ», преподаватель кафедры журналистики Института филологии и журналистики.

А. Г. Солодовников – региональный директор блока технологий «СБЕР», преподаватель кафедры журналистики Института филологии и журналистики Института филологии и журналистики.

И. В. Померанцев – руководитель LM-направления компании GLOBUS IT, преподаватель кафедры журналистики Института филологии и журналистики Института филологии и журналистики

Information about the authors

L. S. Makarova – Candidate of Philology, Associate Professor of the Department of Journalism, Institute of the Philology and Journalism, Deputy Director for Methodological Work.

Y. V. Batashev – System Analyst at BI-TELECOM, Lecturer of the Department of Journalism, Institute of the Philology and Journalism.

A. G. Solodovnikov – Regional Director of the SBER Technology Block, Lecturer of the Department of Journalism, Institute of the Philology and Journalism.

I. V. Pomerantsev – Head of the ML department at GLOBUS-IT, Lecturer of the Department of Journalism, Institute of the Philology and Journalism/

Статья поступила в редакцию 27.06.2024; одобрена после рецензирования 09.07.2024;
принята к публикации 02.08.2024.

The article was submitted 27.06.2024; approved after reviewing 09.07.2024;
accepted for publication 02.08.2024.

Авторы заявляют об отсутствии конфликта интересов.

The authors declare no conflict of interests.