DOI 10.31483/r-113839

Миронова Наталия Геннадьевна

ТЕХНОЛОГИИ МЕДИАБЕЗОПАСНОСТИ: МЕТОДЫ ПРОТИВОДЕЙСТВИЯ ФЕЙК-КОНТЕНТУ В ЦИФРОВЫХ МЕДИА

***Аннотация**: цифровые технологии позволяют автоматизировать процессы дезинформации и информационного противоборства, запускать фейки оперативно и масштабно, организовывать информационные интервенции для переключения внимания населения с одной информационной повестки на другую, для проведения информационных и психологических операций, оказывать определенное воздействие на сознание и поведение целевых групп. Фейки циркулируют не только в поп-медиа и каналах пропаганды, но и проникают в публицистику, в образовательный контент, в научный дискурс, причем, не только как предмет научных исследований, но и как результат фальсификации; сформировался феномен фейковой журналистики и публицистики, пренебрегающей высокими стандартами и принципами профессии. В плотном потоке быстро обновляющейся информации затруднено внимательное восприятие, осложнено различение правды и фальсификации, получение достоверной информации о событиях; общество испытывает беспрецедентное давление дезинформации со стороны цифровых медиа, подрывающей доверие к социальным институтам и каналам социальной коммуникации. Эти обстоятельства актуализируют проблематику совершенствования методологии и инструментов верификации информации и распознавания фейк-контента в сетевых медиа. Цель исследования заключается в анализе признаков фейковости медийного контента и совершенствовании методов противодействия дезинформации в медиапространстве. Результаты: исследованы цели и мотивы создателей и распространителей фейков, признаки и технологии фальсификации цифрового контента, выполнен анализ подходов к борьбе с фейками и фальсификацией информации в сетевых медиа, оценка эффективности различных технологий*

выявления фейк-контента и противодействия его влиянию. Изложена концепция комплексной системы противодействия фейкам в цифровом пространстве. Один из выводов исследования заключается в том, что противодействие влиянию фейк-контента сетевых медиа не сводится лишь к техническим и регуляторным решениям, поскольку механизмы этого влияния социально-психологические; когнитивные науки способны дать методологию эффективного противостояния обману и манипуляции. Исследование адресовано специалистам по медиабезопасности.

***Ключевые слова**: медиабезопасность, фейк, дипфейк, искусственный интеллект, генеративные нейросети, детекторы нейросетевого контента, методы противодействия дезинформации, форензика.*

***Abstract**: digital technologies make it possible to automate the processes of disinformation and information warfare, launch fakes promptly and on a large scale, organize information interventions to shift public attention from one information agenda to another, to conduct information and psychological operations, and to have a certain impact on the consciousness and behavior of target groups. Fakes circulate not only in pop media and propaganda channels, but also penetrate into journalism, educational content, and scientific discourse, not only as a subject of scientific research, but also as a result of falsification; the phenomenon of fake journalism and journalism that ignores the high standards and principles of the profession has emerged. In the dense flow of rapidly updating information, attentive perception is difficult, it is difficult to distinguish between truth and falsification, and to obtain reliable information about events; Society is under unprecedented pressure from digital media disinformation, undermining trust in social institutions and social communication channels. These circumstances actualize the issues of improving the methodology and tools for verifying information and recognizing fake content in online media. The purpose of the study is to analyze the signs of fake media content and improve methods of countering disinformation in the media space. Results: the goals and motives of fake creators and distributors, signs and technologies of digital content*

falsification are investigated, approaches to combating fakes and falsification of information in online media are analyzed, the effectiveness of various technologies for detecting fake content and countering its influence is assessed. The concept of a comprehensive system for countering fakes in the digital space is outlined. One of the conclusions of the study is that countering the influence of fake online media content is not limited to technical and regulatory solutions, since the mechanisms of this influence are socio-psychological; cognitive sciences can provide a methodology for effectively countering deception and manipulation. The study is aimed at media security specialists.

***Keywords**: media security, fake, deepfake, artificial intelligence, generative neural networks, neural network content detectors, methods of countering disinformation, and forensics.*

Введение

Дезинформация с развитием цифровых СМИ стала серьезной проблемой; для многих медиа ресурсов значимой задачей их деятельности становится не достоверное отражение событий действительности, а влияние на умонастроение и поведение целевой аудитории через манипуляцию эмоциями, стереотипами, а важным критерием успеха становится не уровень доверия читателя, а цифровые рейтинги (обеспечиваемые техническими методами). Простота изготовления фейк-контента для мошенничества и манипуляций поведением населения растет по мере развития технологий фальсификации, а стандарты правды в цифровых каналах снижаются. Скорость, с которой создается новостной и аналитический контент, возросла, меняются подходы к производству и продвижению информации, испытывают изменение профессиональные ценности и стандарты журналистики, публицистики под влиянием практик блогинга, социальных сетей, в конкурентной борьбе за читателей и рейтинги. Фейк новости, сенсации фабрикуются нередко ради повышения популярности и посещаемости медиа-ресурсов. В эпоху цифровых медиа фальсификации становятся широкой практикой в сетевых СМИ (порядка 20% новостей – фейковые), подрывая

доверие к СМИ и к социальной коммуникации. Технологии производства информации, основанные на больших языковых моделях машинного обучения, все шире применяются для наполнения контентом новостных, аналитических, корпоративных сайтов и блогов в социальных сетях, для персонализированных пропагандистских кампаний и операций влияния, взаимодействуя с пользователями сложными способами (самые распространенные тактики злонамеренного использования искусственного интеллекта последних месяцев и лет – генерация речи или облика общественных деятелей, фальсификация доказательств в виде изображения вымышленных событий. Дипфейки становятся мощным инструментом широкомасштабных психологических операций, пропаганды и информационных кампаний, применяются для дискредитации, мошенничества – когда дипфейки в видео-звонках и трансляциях используются для обмана граждан, банков (подмена биометрии клиентов). По данным Sumsub, в 2023 году количество дипфейков в мире увеличилось в 10 раз по сравнению с 2022 г. (за этот период в Северной Америке зафиксирован рост их количества на 1740%). В российском сегменте интернет за 3 последние квартала 2023 года количество ресурсов с дипфейками выросло на 1000%; увеличилась и выявляемость фейк-контента.

Феноменология фейк-контента представлена разными уровнями исследования (онтологический, аксиологический, методологический, социально-психологический, а также аспект кибербезпасности). На онтологическом уровне явление исследуется как срез цифровой культуры, как аспект феноменологии лжи или постправды (в работах Д. Робертса, С. Фуллера, С.В. Чугрова (Чугров, 2017), Ю.М. Ершова (Ершов, 2018), А.А. Мовчана, других авторов). Многие исследователи, например, Ю. Хабермас, констатируют рост объема и влияния фейк-контента, трансформацию СМИ под влиянием цифровых технологий (Habermas, 2022) и связанные с этим ценностные и концептуальные изменения медиапространства (Грачев & Евстифеев, 2020); о цифровой трансформации культуры писали Н. Б. Кириллова, (Кириллова, 2023), Л.В. Щеглова (Shcheglova, 2021) и др. Некоторые авторы, – например, К.

Шрайбер в работе «Честная ложь: почему мы продолжаем верить в то, что портит жизнь», – обосновывают эволюционную необходимость лжи с позиции когнитивистики, объясняя доминирование фальсифицированной информации в коммуникативном пространстве. Социально-психологический и методологический аспекты фейка, как инструмента влияния на убеждения и социальное действие, исследовали многие авторы (В.П. Черкасов, Г. Почепцов, В.П. Шейнов, С. Кара-Мурза, И.М. Дзялошинский и др.). Были разработаны модели распространения фейк-контента в сетях (в работах Д.А. Губанова, Н.М. Радько, Е.А. Шварцкопф и др. описаны различные т.н. эпидемиологические и другие модели распространения фейков); исследованы каналы распространения фейков (Ушкин, 2024), способы противодействия влиянию фейк-контенту; технологии создания фейк-контента разного типа; подняты проблемы законодательного регулирования фейк-контента и др. В последние годы уделяется значительное внимание проблемам кибербезопасности и методам противодействия дезинформации с использованием фейков (в т. ч. дипфейков). Исследование посвящено аспектам медиабезопасности, связанным с выявлением фейков в цифровых медиа и способам противодействия негативному влиянию фейк-контента.

Результаты

Фейк представляет собой намеренную фальсификацию информации. Мотивы авторов фейков не всегда очевидны, но, если в сообщении прослеживается стремление автора ввести аудиторию в заблуждение, создать определенный настрой или побудить к определенному действию, есть признаки фальсификации, то сообщение – фейк, а не «честное» заблуждение автора. Фейк-новости после многократного ретранслирования медийными ресурсами начинают жить своей медиа-жизнью как свидетельства «альтернативной правды» (оставаясь симулякрами и фактоидами). Фейк-контент различается по мотивации изготовителя фейк-контента, цели фальсификации: помимо дезинформации, фейки могут развлекать и мистифицировать, поддерживая коммуникативную игру с читателем, отдающим себе отчет в мистификации;

фейки используются и как инструмент персонифицированной дискредитации, травли, троллинга, обмана (например, фейковые видео-звонки жертвам мошенничества). Один из трендов – использование фальсифицированных «научных» публикаций (Pluckrose, Lindsay & Boghossian, 2018), а также статей якобы от имени известных ученых, рассылаемые в новостные агентства и СМИ, чтобы обеспечить текстовым фейковым сообщениям больше доверия, цитируемости либо продвинуть определенную повестку в обществе или целевой группе. Разнообразие мотивов фейкеров породило спектр «жанров» фейков (пародия или сатира, развлекательный контент, провокация и хайп, политический пиар и реклама, пропаганда, имитация (клон) реального лица или ресурса, манипуляция общественным мнением и т. д.). Для эффективного противодействия фальсификации целесообразно формализовать признаки, «маркеры» фейк-сообщений, указывающие на мотивацию создателей и распространителей фейков, позволяющие различать фейки по целевому их назначению. Целями фейк-контента могут быть: (1) продвижение ресурса, продукта, персоны (реклама); (2) переключение и отвлечения внимания в медиапространстве от определенных событий, смещение акцентов в оценке событий, персон в общественном сознании; (3) развлекательные цели (шоу-бизнес и кинематограф, интернет-фольклор, троллинг); зачастую развлекательные интернет-паблики являются частью системы («воронки») вовлечения пользователей в закрытые сообщества деструктивной тематики для идеологической и психологическую обработки; (4) пропаганда, создание нужного умонастроения или общественного резонанса, провоцирование индивидуального или социального действия; (5) кибершпионаж, криминальные и мошеннические цели (кража конфиденциальной информации и финансовых средств) с использованием видео-звонков с дипфеками лиц, голосов; мошенники пытаются получить кредиты банков от имени жертв обмана, пройти дистанционное собеседование для трудоустройства в IT-компаниях в качестве

удаленного сотрудника с помощью дипфейков реальных специалистов[1] и т. п. [6]. Еще одна цель фейков – дискредитация, пранк, травля, манипуляции поведением жертвы (например, политические фейки как средство «черного пиара», порно-фейки и т. п.).

Фейки различны по убедительности и влиянию на «цель», в зависимости от технологии[2] и качества исполнения. Фальсифицированный контент может быть текстовым сообщением, изображением, видеороликом, поддельным биометрическим идентификатором цифровой личности (речь, лицо, поведенческие характеристики и т. д.), клоном/имитацией чужого сайта или аккаунта. Видео (в т.ч. дипфейк) убедительнее текстового фейк-сообщения, поскольку первая сигнальная система (зрение и слух) доминирует над второй (речевой), визуализация сюжета кажется более непосредственным свидетельством, чем его текстовое описание. Видео-фейк воспринимается менее критично, чем текст и фото-фейк, ибо несет больше информации, сильнее задействует внимание, визуальный образ легче вовлекает в сюжет и запоминается. Технически фейк может быть изготовлен «с нуля» (выдуманное событие, созданное GNN-инструментом изображение несуществующего объекта/персонажа без реального прообраза; постановочное фото или видео); либо фейк создается путем манипуляций с объектом-оригиналом (редактированием фото, видео, звукозаписи, клонированием web-страницы, подделкой аккаунта, клон-фишингом документа) или синтезом нескольких реальных изображений (наложение другого фона, маски одного человека на видео-изображение другого (дипфейки), синтез речи голосом определенной персоны, не произносившей этих слов в реальности и т. д.). Фейк ориентирован

[1] В случае успешного найма мошенники рассчитывают получить привилегированный доступ к корпоративным ресурсам, выявить и использовать уязвимости в системе безопасности для кибератак.

[2] Для изготовления фейков применяется ручной фотомонтаж или программно-технические средства автоматизации на основе технологий машинного обучения.

на определенную целевую аудиторию, зачастую усилен приемами социальной инженерии, технологиями пропаганды, SMM-маркетинга, кибер-хакинга[3] и т. д.

Фейки циркулируют в разных каналах социальной коммуникации, и влияние фейк-контента на коллективное сознание зависит от уровня доверия этим каналам. Телевидение, другие традиционные СМИ пользуются относительно высоким доверием, новые СМИ (соцсети, наиболее подверженные фейк-контенту) – низким; соцсети и мессенджеры используют как основной источник информации более половины граждан[4], интернет-версии СМИ – около 30%; непосредственные социальные коммуникации («сарафанное радио») – примерно столько же. При этом «авторитетные» СМИ перенимают у блогеров и цифровых медиа их методы и практики привлечения и удержания внимания аудитории, используют SMM-методы (Social Media Marketing) для продвижения своей повестки, повышения рейтинга и поддержания вовлеченности целевой аудитории (с учетом демографических, психологических, поведенческих особенностей этой аудитории). Большинство традиционных СМИ имеют цифровой аналог (сайт, аккаунт, канал в соцсетях) или используют сюжеты и новости из соцсетей, популярных мессенджеров, собственные бот-каналов (куда любой желающий может вбросить фейк-контент). В условиях конкуренции за аудиторию (и доход) цифровые медиа склонны использовать не всегда этичные, но эффективные приемы привлечения и удержания внимания: броские (в т.ч. гиперболизированные, двусмысленные, пугающие и обманывающие) заголовки, задевающие эмоции анонсы, трагический или пугающий и т. п. контент; лонгриды, заставляющие читателя вовлекаться (и проникаться идеологией или авторским посылом); ссылки-переходы на статьи подобной тематики удерживают читателя на ресурсе; речевые обороты-триггеры,

[3] Например, фейк вбрасывается в медиа-среду от имени «лидеров мнений», через взломанный или клонированный аккаунт, для разгона фейк-сюжета привлекаются бот-фермы комментаторов, лайкеров, ретрансляторов для повышения рейтинга фейк-сообщения в поисковой выдаче соцсетей и браузеров.

[4] Дезинформация стала глобальной проблемой. Радиоспектр [Электронный ресурс]. – Режим доступа: https://rspectr.com/novosti/dezinformacziya-stala-globalnoj-problemoj (дата публикации: 07.11.2023).

задевающие «за живое»; броские иллюстрации (в т.ч. заведомо фейковые, сгенерированные нейросетями) и видеоролики производят более глубокое впечатление, получают больше оценок и репостов, чем текст. Под новостным сообщением администрацией ресурсов организуются дискуссии, чтобы заставить читателя вовлечься, отреагировать, занять позицию; при этом фейк-контент вызывает более бурное обсуждение, чем нейтральное и более объективное новостное сообщение. Публикации на актуальные темы привлекательны, и, чтобы сделать тему актуальной, ее делают вездесущей методами вирусной рекламы и SEO (Search Engine Optimization) для раскрутки. В сочетании с алгоритмами работы поисковых систем, формирующие результаты поисковой выдачи, подобные методы обеспечивают популярность одних ресурсов и точек зрения (понижая поисковую выдачу альтернативной тематики или позиции). Даже сомневаясь в достоверности новости, люди делятся фейк-новостями чаще, чем правдивыми новостями (Vosoughi, Roy & Aral, 2018), не стремясь или не умея их проверить. Фейк-новости до своего разоблачения «живут» хоть и недолго (4–5 дней[5]), но успевают «завируситься» и ретранслироваться, оказав свое убеждающее (и побуждающее) воздействие; при этом и после своего разоблачения они продолжают влиять на аудиторию, став частью социального опыта и памяти. Эффекты отложенного влияния разоблаченного фейк-контента исследованы в 2019 г. специалистами Университета Корка в 2019 году (Fitch, 2019): как оказалось, люди помнят о событиях прошлого (не помня источник новости), черпая воспоминания и из фейковых новостей, т.о., дезинформация оказывает влияние на выбор (ценностный, житейский, политический) и в дальнейшем (следовательно, важнее не допускать появления и распространения фейков, чем заниматься их разоблачением).

[5] Сколько живет фейк в интернете и как его выявить, рассказали эксперты в ходе «дня без фейков». Диалог-инфо [Электронный ресурс]. – Режим доступа: https://dialog.info/skolko-zhivet-fejk-v-internete-i-kak-ego-vyyavit-rasskazali-eksperty-v-hode-dnya-bez-fejkov (дата обращения: 15.01.2024).

Фейк зачастую направлен на изменение убеждений и поведение целевой аудитории фейка. Действенным фейк в конечном счете делает доверие к нему, обусловленное когнитивным состоянием, в которое «качественный» фейк погружает читателя или зрителя (задействуя методы психологии и когнитивных наук[6], социальной инженерии). Чтобы убедить, сообщение должно привлечь внимание, вызывать доверие читателя формой подачи и содержанием, быть понятым и принятым. Фейк-сообщению нужно притвориться в чем-то «своим» или вызвать сильные эмоции, чтобы побудить себя транслировать (и репостить), влиять на мнение или поведение. Доверие вызывает авторитетность (или популярность) медиа-источника, присутствие в сообщении узнаваемых языковых «маркеров» и слов-триггеров, делающих поднимаемую проблему актуальной для читателя, текст – понятным, описываемые события – узнаваемыми и вызывающими эмоциональный отклик, ощущение причастности. Чтобы усиливать доверие, выполнить подстройку к читателю, алгоритмы медиаплатформ и соцсетей отслеживают историю посещений и переходов, интересы и персональные характеристики пользователей, предлагая то, что привлечет внимание, удержит на ресурсе дольше; перепосты в ленте, выдача схожего по тематике контента создают эхокамеры, погружающие в определенную тематику, эмоциональное и ментальное состояние, закрепляя убеждающий эффект (повторение убеждает сильнее однократного сообщения, а непрямое сообщение «между прочим» зачастую убедительное, чем прямо внушающее сообщение и т. д.). Словом, соцсети манипулируют поведением сетевых аудиторий, воздействуя на убеждения, смещая «окна Овертона» в оценке событий, персон, концепций, социальных норм. Фейк-контент, часто встречающийся в социальных сетях, – инструмент подобных «машин убеждения».

[6] В когнитивной психологии, в рамках нейролингвистического программирования (НЛП) разработаны подходы к созданию и изменению убеждений, широко применяемые создателями фейк-контента.

Все чаще не журналисты и эксперты (руководствуясь профессиональным стандартам), а ИИ-боты наполняют контентом новостные, аналитические ресурсы, блоги соцсетей, другие медиа; по оценке АНО «Диалог регионы»[7], количество контента, созданного при участии генеративных нейросетей, выросло в 17 раз в 2023 году по сравнению с 2022 и продолжает кратно увеличиваться. Все чаще для иллюстрации постов и сообщений редакторы и авторы используют не реальные фото- и видеоматериалы, а стоковые и изображения, сгенерированные ИИ-сервисами на основе GAN-моделей (спрос на графический дизайнеров и иллюстраторов упал на 17% за последние 2 года). ИИ-инструменты (ChatGPT, BingChat, Claude, Bard и т. п.) генерируют новостные, технические тексты, обзоры и объяснения, резюме и экспертные заключения[8], код приложений и веб-сайтов, учебных, научных и публицистических текстов, рецензий (Lukeš et al., 2023). Хотя условия сервиса ChatGPT запрещают его злонамеренное использование, он нашел применение и у фейкеров и кибермошенников (для генерации фишинговых писем, вредоносов, выдуманных сообщений, убедительных по форме). Генеративные инструменты на основе больших языковых нейросетевых LLM-моделей оказались склонны (в большей или меньшей степени) к выдаче правдоподобных, но недостоверных или ложных сообщений[9] (их склонность лгать политкорректно называют галлюцинированием[10]); Эта предпосылка недостоверности ИИ-контента усугубляется рядом других слабостей LLM-моделей: статистическим и вероятностным «механизмом» генерации ИИ-контента; отсутствием внутренней логики (ИИ-модель не рассуждает как человек, когда генерирует текст, а

[7] Объем созданного при помощи нейросетей контента за год вырос в 17 раз [Электронный ресурс]. – Режим доступа: https://www.kommersant.ru/doc/6806550 (дата обращения: 26.10.2023).

[8] Например, чат-бот Med-PaLM от Google Research и DeepMind консультирует по вопросам медицинской диагностики (92,6% ответов на открытые вопросы диагностики заболеваний Med-PaLM дала на уровне контрольной группы врачей).

[9] Introducing ChatGPT [Electronic resource]. – Access mode: https://openai.com/blog/chatgpt (30.10.2022).

[10] Например, инструменты для юридической практики, созданные LexisNexis (Lexis+ AI) и Thomson Reuters (Westlaw AI-Assisted Research и Ask Practical Law AI), «галлюцинируют от 17% до 33% времени» (Magesh, Surani, Dahl, Suzgun, Manning & Ho, 2024).

угадывает следующий высоковероятный токен фразы); неспособностью этих моделей видеть без подсказки свои ошибки и выполнять в рамках одного сеанса связанные задачи без специального предложения это проделать; отсутствием прямого доступа к массиву своих обучающих данных[11] (отчего LLM-модель не может привести источников и оснований своих синтетических «рассуждений», а выдумывает их) и к сложным вычислениям (отчего и расчеты часто содержат ошибки); ограничением на размер контекста[12], который учитывается в ходе диалога с пользователем и составляет контекстную память модели в рамках одного сеанса «общения» и др. Поскольку подобные языковые и мультимодальные нейросетевые модели не имеют внутренних «этических» ограничителей, их заменяет внешняя цензура сервиса в виде ряда запретов, фильтрующих запросы пользователей и выходной контент модели. Например, ИИ-сервис Midjourney блокирует некоторые текстовые запросы к модели, а несколько десятков модераторов контента постоянно контролируют использование платформы. Это не мешает использовать сервисы на базе LLM-моделей для создания вредоносного, дезинформирующего и токсичного контента: пользователи выявляют и обходят фильтры запретов нейросети, используют обфускацию (запутывание) текста запросов, ищут и применяют джейлбрейки (например, DAN для ChatGPT), отключающие цензуру; организуют «атаки отравления обучающих данных», чтобы заставить нейросеть выдавать контент определенного характера. В частности, показано (Wei et al., 2022), в результате атак отравления обучающих данных GNN-модели демонстрируют усиление склонности к генерации токсичного, противоправного, недостоверного контента. Например, в 2022 году разработчики представили большую языковую модель Galactica «на благо научного сообщества» (по их словам, превосходящую GPT-3, Chinchilla и PaLM), «способную рассуждать о

[11] Подключение LLM (например, есть у Bing Chat) к поисковой дополненной генерации (RAG) снижает неточность и неактуальность ответов (из-за отсечения знаний периодом до обучения), но поскольку модель та же, она все равно генерирует галлюцинации.

[12] Остается актуальной тема влияния увеличения длины контекстного окна на снижение склонности LLM-моделей галлюцинировать [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2307.03172

научных знаниях», «резюмировать академические статьи, решать математические задачи, создавать статьи для Wiki, писать научный код, аннотировать молекулы и белки и многое другое» (Taylor et al., 2022); однако им пришлось закрыть тестовый доступ пользователям уже через 2 дня после открытия, поскольку модель оказалась способна к генерации убедительных, но совершенно фейковых, текстов, стилистически отвечающих стандартам научной публицистики (конспирологических теорий с выдуманными событиями, фейковых химических формул, трактатов с опасными для жизни рекомендациями (например, о пользе добавления в пищу измельченного стекла) с лживыми отсылками к «экспериментам», сфабрикованными «цитатами» и т. п. [13] Зная подобные неустранимые ограничения своих LLM-моделей, их создатели, рекомендуют относиться к сгенерированной информации своих сервисов лишь как к «гипотезе», нуждающейся в верификации человеком.

Мультимодальные нейросетевые модели способны создавать убедительные и реалистичные иллюстрации, облегчая работу фейк-журналистов и мошенников. Одна из технологий фальсификации – дипфейк, – основана на методах состязательного обучения 2-х нейросетей (GAN, генеративной и классификатора[14]). По подобному принципу создаются и видео-дипфейки (для этой цели создано в последние годы множество инструментов, в т.ч. Doublicat и Reface, DeepFaceLab в Google Colab, Fake-App, Zao для Android и iOS, MachineTube; сервис-фильтр Avatarify Work накладывают дипфейк-фильтры в

[13] Примеры созданнного контента были размещены пользователями сервиса в твиттере (https://twitter.com/mrgreene1977/status/1592958921026985990; https://twitter.com/mrgreene1977/status/1593278664161996801 и т.п.)

[14] Генеративная (G) сеть создает изображения (разворачивая целевое изображение на базе скрытых параметров, синтезируя с ним другие изображения обучающего набора), а дискриминативная (D) нейросеть отсеивает сгенерированные части контента, не похожие на реальные, возвращая на вход G-нейросети результат оценки подделки. По завершении состязательного обучения система способна генерировать правдоподобный фейковый контент определенной тематики. Для изготовления видеофейков (с наложением маски человека-цели на изображения «актера») с помощью нейросети извлекаются кадры изображения цели (чья маска используется в подделке) - и кадры видеоряда с движениями «актера», на чье лицо будет накладываться покадровая маска цели, детектируются контуры заменяемого лица; в ходе обучения G-нейросеть синтезирует оба изображения, а нейросеть-дискриминатор оценивает сходство фейк-кадров с маской; из прошедших отбор фейк-кадров создается фейк-видео.

режиме реального времени во время видеозвонков, видеоконференций; бота Smile Vector может изменить выражение лица на готовом фото; алгоритмы, подобные Face2Face, могут заменить мимику лицу известного человека на мимику актера на видео, динамически поместить человека или персонаж в любой ландшафт или обстановку). Сервисы и программы клонирования голоса для синхронного озвучивания фейк-видео также разнообразны (VeraVoice, Neural Voice Puppetry и др.). Дипфейк-инструменты пользуются популярностью у пранкеров и мошенников, поскольку не всегда требуют даже развитых навыков программирования (хотя качество онлайн фейк-трансляции зависит от производительности видеокарты); образцы изображений и рисунка голоса жертвы создатели дипфейков берут из открытых источников, соцсетей, путем взлома аккаунтов месенджеров с голосовыми и видео-сообщениями. Массовым явлением стали дипфейк-атаки на корпоративные группы в месенджерах; дипфейки руководителей организаций используются для обмана их подчиненных и коллег), дипфейк-стримы от лица известных блоггеров – для обмана подписчиков; цель дипфейк-атак – получения от жертв обмана денег или конфиденциальных сведений.

Умение распознавать признаки фейков – полезный навык не только для специалистов по кибербезопасности или журналистов, но и для широкого круга потребителей медиаконтента, поскольку фейки вездесущи. На возможную фейковость текстового сообщения указывает его тональность и стилистика: кликбейтность заголовка, эмоциональная накачка в тексте сообщения; слова-триггеры, «цепляющие» внимание, «подстегивающие» эмоции (страх, гнев); структура сообщения построена так, чтобы вовлекать и удерживать читателя (чем дольше действует вовлеченность в сообщение, тем глубже сформированное им убеждение). К признакам фейковости сообщения можно отнести отсутствие авторства сообщения (либо имя вымышленное и, не «гуглится»); упоминаемые в фейк-сообщении факты описаны так, что их невозможно или сложно проверить (например, приводится свидетельство никому не известного человека или источника, цитаты – без ссылок на источники либо источники сомнительны).

Для фейк-сообщений нередко характерно низкое качестве изложения (шаблонные, пропагандистские, рекламные или просторечные речевые обороты и выражения-триггеры, манипулятивные приемы и алогичность и т. п.), иллюстрации сгенерированы нейросетью или взяты из фото-стока, без указания авторства и т. п. Фейки обычно публикуются изначально на «серых» ресурсах интернета, в редких доменных зонах, на сайтах, не имеющих сертификата безопасности, а оттуда разносятся ботами по социальным сетям и мессенджерам, откуда их зачастую черпают и цитируют более авторитетные медиаресурсы и т. д.

Часть недостоверного цифрового контента производят большие языковые модели генеративных неросетей (склонные галлюцинировать), затем он транслируется в других медиа. Признаками текста, созданного нейросетевым сервисом, пока могут служить стилистические, логические ошибки, несогласованности текста; термины на протяжения текста могут использоваться то в одном, то совсем в ином контексте, создавая смысловые нестыковки между частями текста; могут присутствовать выдуманные названия, цитаты несуществующих экспертов (что выявляется даже поверхностной проверкой). Впрочем, если сгенерированный текст отредактирован человеком, эти признаки могут и отсутствовать.

Признаки дипфейков: отсутствие у изображения метаданных, автора; фото-дипфейк выглядит слишком гладко (нет зернистости реальных фотографий) и идеализированно; дипфейковые (и вообще созданные GAN-нейросетями) изображения плохо передают отражения в зеркалах, окнах, глазах (оно отсутствует или неестественное), не бывает эффекта «красных глаз» (в отличие от некоторых реальных фотографий) и т. п. На видеофейке может присутствовать мерцание и искривление контура лица (если «наложена» цифровая маска на кадры видео), нечеткость волос, зубов и других мелких деталей, скачки качества видео в разные моменты; внутрикадровые и временные несоответствия; неестественность темпа дыхания и моргания, движения глаз (Jung et al., 2020), мимических мускулов, других спонтанных физиологических

проявлений на видео (впрочем, алгоритмы подделки совершенствуются и в этом направлении); речь дипфейка может звучать слишком чисто, равномерно и без пауз, без звука дыхания, без слов-паразитов и междометий; неестественны модуляции голоса, неправильны ударения в словах и т. п. Если фейк используется для обмана в удаленной цифровой коммуникации зачастую мошенники использует фейк-запись, а не дипфейк в реальном времени (требующий больше ресурсов или навыков), так что отсутствие возможности обратной связи (или спонтанности видео-чата) должно насторожить.

«Поведенческими» признаками фейк-контента могут служить агрессивность раскрутки в сетевых медиа (мессенджерах, соцсетях, форумах); использование ботов, фабрик троллей, устраивающих «дискуссии» в комментариях к фейк-сообщению, лайкающих и репостящих фейк для повышения его рейтинга и упоминаемости в сетевых медиа (альтернативные же ресурсы и точки зрения подвергаются замусориванию ботами-спамерами). Люди, склонные к конформизму, видя кликбейтную «новость», под которой множество лайков и комментариев, воспринимает ее заслуживающей внимания и репоста. При этом сложно понять, кто ведет канал или комментирует новость (фейковый аккаунт-бот или человек), поскольку боты-комментаторы и бот-каналы имитируют поведение людей. Признаками фейк-аккаунта, фейк-канала, созданного для фейковых вбросов и манипулирования общественным мнением, могут служить высокая активность при короткой истории существования аккаунта, сотни и тысячи подписок акаунта на другие каналы, отсутствие или минимум взаимных друзей (у реальных аккаунтов-людей число подписчиков (друзей) обычно от 20 до 200–300), специфический контент. Читатели нередко понимают, что новостное или аналитическое сообщение создано не человеком, а нейросетью, и не склонны доверять ему; «52% респондентов из США, 63% опрошенных из Великобритании, Южной Африке – 81% сообщили, что им было бы некомфортно читать новости, созданные с помощью нейросетей» (Newman, 2024); треть россиян отказались бы читать медиа, в котором тексты пишутся при помощи нейросетей, по оценке Rambler&Co.

Впрочем, GAN-модели применяются не только для изготовления реалистичных дипфейков, так и для их обнаружения (Marra, Gragnaniello & Verdoliva, 2018). Программно-технический инструментарий распознавания фейков развивается, помогая определять, подвергалось ли изображение фальсификации, не является ли контент синтетическим ИИ-продуктом, но широкому кругу пользователей сетевых медиа этот инструментарий малодоступен (программы и сервисы платны, требуют профессиональных IT-навыков, доступ к ним зачастую запрещен для россиян и т. д.).

Дезинформация в цифровых медиа распространяется и действует подобно заражению вирусом, поэтому диапазон мер и технологий противодействия информационному «вирусу» подобен спектру мер борьбы с вирусными заболеваниями (от профилактических и предупредительных – до терапевтических). Для «профилактика», снижающая риск «поражения» дезинформацией, необходимы законодательные и организационные меры регулирования[15] деятельности масс-медиа, стандартизация и регламентация разработки и применения ИИ-инструментов (облегчающих фейкерам их деятельность), пресечение деятельности медиа-ресурсов и лиц, использующих фейк-контент для обмана и противоправных действий, повышение уровня медиаграмотности в обществе. «Терапевтические» мероприятия состоят в идентификации, удалении (и разоблачении) фейк-контента из медиапространства, донесении до аудитории правдивой информации; создание доступных сервисов, позволяющих проверить достоверность[16] информации или разоблачить фейк. Оба направления борьбы с фейк-контентом (предотвращение и пресечение фальсификации информации) представлены мерами

[15] К слову, большинство людей приветствуют госрегулирование фейков (88% опрошенных граждан 16 стран мира, согласно исследованию Ipsos [Электронный ресурс]. – Режим доступа: https://rspectr.com/novosti/dezinformacziya-stala-globalnoj-problemoj (дата обращения: 07.11.2023).

[16] Сейчас действует ряд фактчекинговых сервисов, в т.ч. российские «войнасфейками.рф», https://lapsha.media (от «Лапша Медиа групп»), зарубежные www.snopes.com, www.factcheck.org, fullfact.org, PolitiFact, «International Factchecking Net-work» (https://azbukamedia.com/category/fakefighter) и др.

организационно-правового и программно-технического характера, и в комплексе формируют арсенал противодействия дезинформации; меры охарактеризованы ниже.

1. Законодательные, регуляторные меры противодействия распространению фейк-контента позволяет упорядочить работу цифровых медиа, вывести деятельность, связанную со злонамеренным использованием фейков, за рамки правового поля, определить признаки такого рода деяний и меры борьбы. Большинство граждан одобряют идею регулирование фейков: 83% опрошенных граждан 16 стран мира, согласно исследованию Ipsos (Ipsos, 2023); 82% россиян в 2023 году одобряли блокировку фейков в интернете как регуляторную меру (ВЦИОМ, 2023). В законах ряда стран есть запреты на использование дипфейк-технологии в определенных целях. Так, в Китае с 2019 г. закон запрещает публикацию фейк-новостей, сгенерированных ИИ-инструментами, использование генеративного ИИ при подаче заявок на финансирование научных исследований; действует требование маркировки ИИ-контента. Законодательство Германия предусматривает ответственность для владельцев web-ресурсов за размещение фейк-контент, несущего общественную угрозу. В ряде американских штатов (Вирджиния, Техас, Калифорния, Тенесси) закон регламентирует применение дипфейк-технологий, в 2024 году Конгресс США подготовил законопроект по маркировке ИИ-дипфейков[17] корпорациями, разрабатывающими генеративные нейросетевые модели. Еврокомиссия собирается обязать IT-компании маркировать контент, созданный их нейросетями. С июня 2024 года и российские законодатели работают над проектом закона о маркировке контента нейросетей. В ряде стран закон также защищает приватность и конфиденциальность граждан от манипуляций, поскольку их биометрические признаки могут быть использованы для обучения нейросетей, подвергаются рискам утечки и дипфейк-подделки. В этом плане США, Канада, Европейский Союз (ЕС), Великобритания, Китай, Япония, Корея,

[17] Eshoo Anna G. Press Release (March 21, 2024) [Electronic resource]. – Access mode: https://eshoo.house.gov/media/press-releases/rep-eshoo-introduces-bipartisan-bill-label-deepfakes

Сингапур применяют риск-ориентированный подход к регулированию ИИ, предполагающий более строгие обязательства по обеспечению безопасности и прозрачности ИИ-технологий с высоким социальным риском (хотя IT-компании лоббируют смягчение[18] законодательного регулирования ИИ). В России также применяется риск-ориентированный подход в регулировании ИИ[19] ; в отношении фейков в России действуют статьи Административного кодекса (например, ст. 13.15 КоАП РФ злоупотребление свободой массовой информации предусматривает наказание за распространение фейк-новостей), статьи Уголовного кодекса (например, ст. 207.1–2 УК РФ – за распространение заведомо ложной информации; закон об уголовной ответственности за фейки о действиях российских военных); ФЗ №31 от 18.03.2019 г. «О внесении изменений в статью 15.3 ФЗ «Об информации, информационных технологиях и о защите информации» о блокировке фейковой информации; в последние месяцы обсуждаются инициативы законодателей с требованием маркировки сгенерированного нейросетями контента с подменой лица, голоса, подменой сцен на видео, изменения в ст. 152.1 ГК РФ о синтезе голоса ИИ-технологией, об охране голоса как части цифрового образа граждан и т. д. Понятие «дипфейк» в законах РФ не определено, хотя отдельные статьи Гражданского и Уголовного кодексов запрещают противоправные деяния с использованием сфальсифицированного контента (например, ст. 159 УК РФ «Мошенничество», поправки в УК РФ об ответственности за распространение фейков о действиях Вооруженных Сил России). В настоящее время поправки в законодательство за отдельные деяния с использованием технологий дипфейков готовят Минцифры, МВД, РКН; Госдума РФ обсуждает изменения уголовного наказания за деяния с

[18] Так, OpenAI настояла на том, чтобы не считать высокорисковыми ИИ-системы, в т.ч. ChatGPT, DALL-E.

[19] П. 51 «Стратегии развития искусственного интеллекта» в России, утвержденный Указом Президента РФ от 10 октября 2019 г. №490 «О развитии искусственного интеллекта в РФ» гласит, что «риск-ориентированный подход: уровень проработки, характер и детализация изменений при регулировании вопросов в области искусственного интеллекта должны соответствовать уровню рисков, создаваемых конкретными технологиями и системами искусственного интеллекта для интересов человека и общества».

использованием дипфейков (статьи 128, 158, 159, 163, 165 УК РФ). Хотя в РФ «ответственность за все последствия работы систем искусственного интеллекта всегда несет физическое или юридическое лицо» (Указ Президента РФ от 10.10.2019 г. №490), – как представляется, в законах должна быть более четко определена доля ответственности за фейк-контент. Пока ответственность неопределенно «размазана» между создателем нейросетевого сервиса (например, ИИ-бота) и теми, кто распространяет или использует фейк как реальное свидетельство. Если фейк-контент собирает и публикует (или генерирует и публикует) ИИ-бот, ответственность за последствия от такого контента (в т.ч. фейкового) должен отчасти нести и разработчик интеллектуального бота, и пользователи бота в блогах и медиаканалах (которые, однако, могут не знать нюансов и последствий технологии), и те, кто ретранслирует фейк-контент. Но разработчик ИИ-бота или сервиса не знает определенно, какой конкретно контент будет сгенерирован в дальнейшем с помощью его продукции. Впрочем, фейк-контент в больших объемах вбрасывают и медиа-ресурсы, находящиеся и вне российской юрисдикции, поэтому нормативно-правовое регулирование лишь ограниченно действенно.

2. Организационные меры противодействия фейк-контенту применяются и на государственном уровне, и бизнесом (владельцами платформ, соцсетей, поисковых сервисов). Например, в Китае Cyberspace Administration of China (CAC) обязала с 2023 года всех владельцев онлайн-сервисов проверять и удалять аккаунты, замеченные в создании или распространении фейков. В России Роскомнадзор практикует блокировку сетевых ресурсов, занимающихся противозаконной деятельностью, замедление и др. технические меры защиты россиян от противоправной деятельности в цифровом пространстве. Владельцы медиа-платформ, нейросетевых инструментов, поисковых систем все чаще добровольно цензурируют, фильтруют, блокируют бот-аккаунты (как «ВКонтакте») по жалобам на фейк-контент и противоправный контента, помечают фейк-контент знаком «ложная информация» (как Twitter (X), Facebook), блокируют фразы, характерные для фейк-контента (как мессенджер

Wechat корпорации Tencent), организуют сообщества по разоблачению фейков и т. д. Владельцы популярных поисковиков блокируют фейк-контент по жалобам пользователей (например, Google удаляет по жалобам пользователей вредоносный контент (например, интимные дипфейки), понижает рейтинг сайтов с большим количеством жалоб на дипфейки, фильтрует поисковую выдачу с именами потенциальных жертв дипфейков). Фотостоковые сервисы, чтобы избежать судебных претензий, фильтруют загружаемый пользователям мультимедиа-контент, рекомендуют помечать размещаемые изображения тегами, позволяющими отличить искусственный контент от реального фото и т. п. (например, Adobe Stock в руководстве пользователей запрещает публикацию дипфейков и изображений, созданных с использованием запросов об определенных людях, местах или объектах собственности, призывает снабжать ключевыми словами отредактированные фото или искусственно сгенерированные изображения). Однако заставить все сетевые ресурсы регулировать деятельность своих пользователей с фейк-контентом технически сложно, к тому же политика фильтрации фейк контента у разных ресурсов сильно различается, поэтому эффективность организационных мер противодействия фейкам невысока и дополняется другими методами. Недостаточная активность силовых органов в отношении кибер-преступников и мошенников также способствует «популярности» фейк-технологий у криминальных групп (в распоряжении которых – обширная «серая» зона медиа-ресурсов, баз и кибер-инструментов из darknet, библиотеки открытого исходного кода ИИ-ботов, генераторы фейковых изображений и видео). Одна из причин активности фейкеров и мошенников, использующих фейк-технологии, – чувство безнаказанности из-за анонимности и недосягаемости в интернете; поэтому ограничение анонимности в цифровом медиа-пространстве представляется полезной мерой для снижения количества фальсификаций и импульсивных репостов фейков. Регулирование должно быть направлено, в идеале, на создание условий, при которых невозможно или рискованно ретранслировать

дезинформацию, фейк-контент, а участники медиа-пространства вынуждены следовать более строгим критериям и ценностям правдивости.

3) Программно-технические меры противодействия фейкам, в т.ч. дипфейкам, заключаются в идентификации и распознавании фейков. Можно выделить 3 подгруппы: (1) методы защиты контента от фальсификации в будущем; (2) методы и технологии маркировки недостоверного контента сразу при его создании; (3) методы детектирования фейк-контента.

3.1. Технологии первой подгруппы позволяют заранее защищать изображения от последующего манипулирования ими (в т.ч. от подделки и создания с их помощью дипфейков). Есть сервисы, которые защищают фото перед публикацией в интернете путем добавления на фото незаметных глазу шумов для защиты от последующего искажения и т. д. В частности, вставка в реальное фото или видео цифровых артефактов, маскирующих те группы пикселей, по которым ориентируются популярные программы распознавания лиц и дипфейк-алгоритмы; при попытке подделать так измененное цифровое фото результат подделки окажется неузнаваем и малополезен для мошенников (например, группа исследователей Бостонского университета предложила метод защиты публикуемых в интернете фотографий и видео от их изменения с помощью дипфейк-технологий: наложение на фото / видео фильтра, отображающего пиксели особым образом; если скачавший защищенное изображение пытается с помощью нейросети создать на его основе дипфейк, изображение становится неузнаваемым и непригодным для обмана). Есть технологии защиты фото-контента от скачивания с использованием CSS свойств, HTML-скрипта и JavaScript-функций (не слишком хорошо защищают от подделки и дальнейшего несанкционированного использования фото), применяемые некоторыми web-сайты. Еще один (слабый) прием защиты изображений на web-страницах от несанкционированного использования состоит в наложение поверх изображения прозрачного пустого слоя (при копировании фото с сайта скопируется только этот пустой слой). Некоторыми разработчики фоторедакторов (Photoshop, Lightroom и др.) предлагают помещать

на защищаемое изображение «водяной знак» для защиты изображений от несанкционированного использования (в т.ч. подделки). Другой способ защиты изображений от дальнейшей подделки заключается в снабжении их метками, которые при попытках редактирования изображения искажаются или удаляются, позволяя понять, что фото – не оригинал, а фальсификат. Роль метки могут служить метаданные (EXIF-метки) и аналоги сертификата цифровой подписи или сертификата доверенного сервера. Подобную технологию использует компания Adobe, разработавшая для своих программ и сервисов функции маркировки крипто-тегами фотографий, редактируемых в Photoshop, что позволяет отслеживать историю изменения фото с момента создания (впрочем, эти теги можно удалить из фото-файла и добавить поддельные). Производители цифровых камер в момент съемки каждое фото и видео снабжают EXIF-метами, что позволяет определить источник контента, когда, где, с какими техническими параметрами сделан снимок. При редактировании в фото и видеоредакторах EXIF-данные автоматически меняются, частично удаляются, – поэтому отсутствие метаданных у изображения намекает на манипуляции с оригинальным фото или видео. Впрочем, EXIF-метки могут быть вручную отредактированы (программами типа EXIF Pilot) и просто удалены (например, фото-стоки и соцсети, где пользователи выкладывают фото для публичного использования, нередко автоматически удаляют EXIF-метки из соображений конфиденциальности[20]), поэтому по скаченному из открытых источников фото или видео вряд ли можно по EXIF-меткам определить, оригинальное или сфальсифицированное это изображение. Еще один метод защиты от фальсификации – размещение доверенного контента в верифицированном по некоторым стандартам (C2PA и т. п.) виде на доверенных хостах (что позволяет, в случае появления в будущем дипфейка, предъявить оригинал и доказать фейковость подделок).

[20] Чтобы не позволять отследить человека или объект по метаданным фотографии тем, кто занимается доксингом, кибершпионажем и т. п.

Изображения, размещенные в цифровых медиа, многократно скачиваются и редактируются, в т.ч. фальсифицируются. Чтобы можно было достоверно определить факт подделки, подмены изображений, они должны иметь более надежную, криптографическую защиту, например, в виде аналога сертификата электронной подписи, проставляемой в каждой отредактированной версии электронного документа; метку-признак изменения изображения должны, в идеале, проставлять автоматически любые фоторедакторы, а история правок и список меток изображений могли бы храниться и накапливаться по принципу технологии блокчейн, позволяющей обеспечить децентрализованную валидацию цифрового контента, отслеживая и подтверждая его источники (примерно так, как NFT-платформы обеспечивают авторство и уникальность NFT-медиафайлам, верифицируя отсутствие аналогов). Сервисы генеративных НС, фоторедакторов, видеоредакторов целесообразно обязать также автоматически маркировать любой отредактированный или сгенерированный мультимедиа-контент уникальной меткой автора/редактора (или записью в реестре блокчейн). У каждого автора или дизайнера в таком случае должна быть уникальный сертификат обязательной электронной подписи, позволяющий подтверждать личность автора цифрового контента при любом редактировании фото или видео. Данный подход мог бы быть действенной защитой от фальсификации контента, если бы стал общеобязательным стандартом для всей отрасли разработчиков инструментов работы с изображениями, но в ближайшее время подобных отраслевых стандартов ожидать не приходится из-за сопротивления всех сторон и технических сложностей. Впрочем, есть отдельные платформы и инструменты верификации цифрового контента, двигающиеся в этом направлении; например, сервис аутентификации фото и видео truepic.com заверяет контент в блокчейнах для верификации цепочки пользователей, работавших с изображениями на разных этапах, от съемки до хранения.

3.2 «Честное предупреждение» – маркировка ИИ-продукции при ее создании, – один из методов борьбы с фейками; целесообразно, чтобы все сервисы, генерирующие нейросетевой контент или выполняющих нейросетевую

обработку фото, видео, звукозаписей, в обязательном порядке ставили на продукт нейросети машиночитаемую (а лучше «человекочитаемую»[21]) пометку «сделано нейросетью». Это позволяет отслеживать фальсифицированный нейросетями контент и отличать его фотофактов и видеофактов. Пока лишь некоторые сервисы используют технологию пометки сгенерированных изображений согласно стандарту IPTC: например, Adobe с Nikon, BBC, Microsoft и Truepic намерены снабжать меткой CR (Content Credentials) изображения, созданные или исправленные с помощью нейросетей (CR-данные будут добавляются в метаданные фото или видеозаписи фоторедактором либо цифровой камерой); сервисы на основе нейросети DeepMind от Google будут вставлять знак Syn-thID в сгенерированные изображения, чтобы распознавать по этому знаку искусственные изображения; в нейросетевом генераторе изображений Vertex AI можно (Gowal & Kohli, 2023) вставлять в сгенерированные этой НС изображения цифровой знак SID. К слову, поисковик Google при поиске изображений может по метаданным изображения указать, что оно сгенерировано нейросетью, если изображение было размечено НС-сервисом.

3.3. Технологии выявления фейк-контента (текста, фото и видео-контента, голосовой звукозаписи, фейковых сайтов и аккаунтов) различны, в зависимости от вида фейка.

Для оценки фейковости текста анализируются: фактология (путем сравнения сообщения с доверенными новостными сайтами и базами), авторы, достоверность ссылок, фото и т. д. Поскольку для фейков зачастую характерна определенная тональность и структура текста и заголовка, признаки манипуляции, логические нарушения, – то контент-анализ помогает оценить вероятность фальсификации. Фактчекинговые сервисы, помогающие провести проверку сообщений на фейковость: российский СКАН (scan-interfax.ru);

[21] На практике такой метод применяется некоторыми разработчиками редакторов, но метки незаметны для человеческого глаза или содержатся в метаданных файла (увидеть которые можно лишь с помощью программных инструментов, часто платных, а не непосредственно), поэтому технология не мешает вводить в заблуждение потребителей новостного контента и изображений.

зарубежные Storyful.com, FactCheck.org, Politi-Fact.com, The Fact Checker (от Washington Post), Mediakritika.by, Trooclick, Truth Goggles, Lazy Truth, Skeptive, Genius и др. Появляются продвинутые (на основе машинного обучения) средства автоматизации фактчекинга. Например, нейросеть Ai Wiz (https://aiwiz.ru/ai_fact_checking) обучена приемам фактчекинга текстов, ссылок, цитат, сопоставляя их с проверенными источниками (в т.ч. с научными реестрами и публикациями), выявляя несоответствия и ошибки; модель для фактчекингового сервиса fakenewsai (https://www.fakenewsai.com, разработчик К. Сингхал) обучена искать признаки сходства проверяемых ресурсов с сайтами фейк-новостей. Фейк-детекторы сетевого контента (с разной степенью эффективности, главным образом на основе технологии искусственного интеллекта): платформа TruthBird (https://www.Truthbird.com) – ищет фейки, руководствуясь своей базой сайтов фейков; BotSlayer (https://www.Botslayers.Com) реализует методы поведенческого анализа аккаунтов пользователей соцсетей для выявления ботов-спамеров; CredEye (https://dl.acm.org/doi/fullHtml/10.1145/3184558.3186967) распознает фейковые сообщения методами семантического анализа эмоциональной окраски текстов, признаков «предвзятости», поиска логических ошибок, проверки по базе достоверных источников; браузерное расширение MediaSifter (https://www.Mediasifter.com) – имеет функцию распознавания дезинформации; Actus (https://actusdigital.com/automatic-ads-detection-and-ai/) – распознает манипулятивные и рекламные тексты. Все анализаторы текстовых фейков имеют определенный процент ложных срабатываний, ограничены обучающим набором, функционалом, тематикой и языком проверяемого контента. Если текст создан нейросетью, его достоверность под вопросом (поскольку всем большим языковым GAN-моделям свойственно «галлюцинирование»), поэтому при проверке фактологии сообщения полезна и проверка на искусственность его происхождения. В детектировании фейк-контента и недостоверного контента, сгенерированного чат-ботами, все чаще применяются технологии машинного обучения. Для распознавания сгенерированного контента используется метод

отпечатков текста, поиск маркеров ИИ-продукции (если эта продукция была маркирована при создании). Для оценки искусственности текстового контента на гитхабе (по адресу https://github.com/HendrikStrobelt/detecting-fake-text) выложен пакет «GLTR: Гигантская тестовая комната для языковых моделей» для определения того, был ли текст сгенерирован генеративной сетью (GPT-2 и т. п.); опубликован инструмент FakeTextDe-tector для распознавания текстов, сгенерированных нейросетями (https://arxiv.org/abs/1805.08751), в т.ч. детектирует фейки и манипулятивные тексты. ИИ-анализаторы текста обычно указывают на подозрительные фразы в анализируемом сообщении, но испытывают сложности с оценкой доли нейросетевого контента в сообщениях или с оценкой вероятности того, что текст был создан нейросетью. Например, детектор ИИ-текста от OpenAI обнаруживает лишь 26% ИИ-контента (Kirchner et al., 2023). При проверке инструментов, распознающих нейросетевой контент («Война с фейками», FactCheck, Politifact, Snopes, «Лапша Медиа», «Лига безопасного интернета», Oigetit Fake News Filter, Alt News, Full Fact, AFP Fact Check) выявлено (Тумбинская & Галиев, 2023), что 80% из них поддерживают проверку фейк-новостей на основании экспертной оценки, остальные 20% основаны на ИИ-технологии), причем, оценка фейковости может быть пристрастной и малополезной для верификации русскоязычного контента (лишь 30% поддерживают проверку русскоязычных новостей). Если в сгенерированный нейросетью текст дорабатывался человеком, способность ИИ-фейк-детекторов распознавать фейк-контент резко падает (например, в случае с контентом ChatBot с 75% до 42% (Weber-Wulff et al., 2023)). Д. Вебер-Вульф с коллегами (Берлинский институт техники и экономики) по результатам оценки 14 инструментов-распознавателей контента, сгенерированного с помощью ChatGPT, заключили, что все эти инструменты с трудом детектировали ИИ-текст, который немного отредактировали люди.

Для обнаружения фейковых изображений применяются методы машинного обучения, программы анализа цифровой подписи и проверки подлинности видео (и технологии, основанные на аналогах блокчейна). Для обнаружения

фальшивых профилей, с которых могут размещаться фейки, применяются ИИ-инструменты поведенческого анализа. К популярным у разработчиков средств форензики методам выявления признаков фальсификации цифровых изображений относятся: анализ метаданных файла изображения; анализ пикселей, поиск аномалий, алгоритмы распознавания признаков фальсификации; ELA (Error Level Analysis) – поиск неоднородностей уровня сжатия изображения; поиск аналогов по сетевым хранилищам изображений и т. п. Сопоставление параметров изображений с реальными событиями, о которых как бы свидетельствует изображение (с учетом EXID-данных файла[22], сетевых источников), позволяет обнаружить признаки подмены (несоответствия освещения, перекраску, монтаж, аномалии изображений или видео). Иногда в EXIF-метаданных изображений в некоторых соцсетях может сохраняться и миниатюра исходного изображения картинки, при редактировании изображений в некоторых фоторедакторах (например, Corel photo-paint X8) оригинал сохраняется и может быть извлечен впоследствии для сравнения с окончательной версией улучшенного (или фальсифицированного) изображения. Отсутствие метаданных может дополнительно свидетельствовать о манипуляциях с изображением, поскольку модификация изображения в фоторедакторов меняет или уничтожает EXIF-данные фотографии, а в сгенерированных нейросетями или скачанных из интернета изображений EXIF-данные изначально отсутствуют; впрочем, метаданные фото могут быть отредактированы или удалены[23] автором (без изменений самого изображения) и из соображений конфиденциальности; да и сами СМИ, некоторые соцсети

[22] Метаданные файла (дата, GPS-метка места сьемки, параметры фокусировки, выдержки, проч.), если не удалены, помогают соотнести фотографию с известными условиями на местности (временем суток, освещённостью, положением теней, окружающий ландшафт в реальности), чтобы оценить, не является ли фотоиллюстрация подлогом. EXIF снимка можно просмотреть в свойствах файла (например, в проводнике ОС Windows), в фоторедакторах (типа Photoshop), сервисах (ACDSee, ShowExif, fakeimagedetector.com), на гитхабе есть код программ для вытаскивания EXIF-данных из фотофайлов и т.д.

[23] EXIF-данные файла можно редактировать и удалять на устройстве, которым сделан снимок, с помощью программ и сервисов (ExifTool, Ghiro, gpsphoto.ru, products.groupdocs.app/metadata); в свойствах файла в проводнике Windows и т. д.

автоматически удаляют метаданные при загрузке фото пользователями ради безопасности. Для обнаружения признаков постобработки изображений используются фильтры/режимы просмотра изображений; анализ пикселей помогает обнаружить признаки манипуляций с изображения в фото-редакторах; ELA-анализ уровня ошибок JPG-изображений показывает различия в уровнях компрессии участков изображения (если различия существенны, это свидетельствует о редактировании изображения (в т.ч. фальфификации); но ELA-метод неэффективен, если фотоизображение многократно скачивалось и сохранялось на носителе, т.к. уровни компрессии участков изображения при этом нивелируются, не позволяя обнаружить признаки редактирования. Применяются и методы анализа пиксельного шума: оригинальные фотографии имеют высокий уровень и однородность шума, а отредактированные в фоторедакторе элементы почти не имеют шума; специальные настройки камеры тоже понижают уровень естественного шума на фото. Применяются и методы машинного зрения для анализа цифровых характеристик изображения, например, анализ глубины пикселей изображения (как это делает нейросетевая модель DepthFake итальянских исследователей), периферийной области и формы лица изображения. Для экспертизы фото-фейков специалисты используют популярные сервисы анализа изображений; FotoForensics[24], JPEGsnoop, Serelay, Truepic, Forensically анализируют яркость, шум, контраст, выявляют клонирование элементов, ELA-фильтр; 29a.ch/photo-forensics поддерживает поиск признаков редактирования и клонирования изображения; сервис Ghiro – поиск признаков монтажа; линейка продуктов класса Digital Fake Prevention (от

[24] В частности, платформа FotoForensics (не позволяющая загружать фотографии из России без VPN) поддерживает не-сколько методов анализа изображений: функция Metadata показывает EXIF-данные изображения; фильтр ELA отмечает белым цветом отредактированные фрагменты (например, вставки, изменения яркости или контраста) – но если откорректированное фото много раз пересохранялось, или изображение является продуктом генеративной сети, то резкие контрасты отсутствуют, зато на ELA-фильтре виден шум в виде красных и синих полос; функция Hidden Pixels показывает белым или черным цветом скрытые пиксели исходного изображения, если был наложен другой слой поверх оригинальных элементов фото (фотошоп окрашивают скрытые пиксели в белый цвет, Gimp и PicMonkey – в черный) и т.д.

Oz Forensics) для определения подлинности цифровых документов и фотографий и т. д. Для дополнительной проверки достоверности фотоизображения может быть выполнен обратный поиск оригиналов изображения (если они когда-либо загружались в интернет) браузерами или сервисами типа tineye.com/search, images.google.com и т. п. Многие сервисы платные и не недоступны из России.

Для проверки фейковости видео применяется покадровый анализ: размытость на отдельных кадрах указывает на возможное редактирование кадров и т. п. Для автоматизации анализа видеофейков применяется машинное обучение; среди ИИ-инструментов анализа – Media Forensics от DARPA; сервис Spotdeepfakes.org поиска признаков дипфейков методами анализа пикселей, параметров сжатия и т. п.; инструмент Angora (от Gfycat, 2019); технология Video Authenticator (от Microsoft, 2020 год). Есть ряд сервисов для распознавания голосовых дипфейков (aivoicedetector.com и др.). В США с 2016 года реализуется проект MediFor (Media Forensics) (DARPA) для оценки достоверности новостного контента, включая фото и видео, разными способами (путем анализа цифровых «следов», метаданных и параметров сжатия изображения, анализа «физических» характеристик изображения (адекватность освещения и теней) и сравнения их с изображенной реальностью (соответствует ли фактический ландшафт, время года, суток, погода – изображению, с учетом метаданных и т. п.). Adobe Research с Калифорнийским университетом в 2019 году разработали инструмент Photoshop Face Aware Liquify для определения фото- и видеомонтажа (и восстановления оригинального вида фото). В 2020 г. ученые из Германии предложили метод машинного обучения на основе частотного анализа с использованием дискретного преобразования Фурье для различения фотографий лиц и дипфейков. На гитхабе (https://github.com/resemble-ai/Resemblyzer) в 2020 году размещен открытый код Resemblyzer инструмента на основе машинного обучения для оценки фейковости видеороликов или голосовых записей и т. д. Microsoft в 2020 г. опубликовала инструмент Video Authenticator для обнаружения дипфейков на основе анализа пикселей на границах предполагаемого наложения дипфейка на реальное изображение.

Корейское web-приложение для Android KaiCatch определяет аномальные искажения лиц (с заявленной точностью 90%). На основе датасета от Google и Jigsaw из нескольких алгоритмов для подмены лиц создана модель (и детектор дипфейков, созданных с помощью нейросети StyleGAN) Assembler, обнаруживающая специфичные манипуляции с изображением (например цветовые несоответствия и аномалии, фотомонтаж, клонированные объекты на фото); модель тестировали фактчекинговые СМИ (Agence France-Presse, Animal Politico, Rappler). Известны программа FakeBuster для выявления дипфейков во время трансляций в Zoom и Skype, сервис Deepware (scanner.deepware.ai) для обнаружения дипфейковых видео. Заметно стремление государств проектировать собственные платформы фейк-детекторы, в т.ч. мультимодальные. Например, в США в 2019–2024 гг. Агентство DARPA (США) инициировало в целях «противодействия атакам дезинформации» разработку программно-аппаратного комплекса Semantic Forensics (SemaFor, https://semanticforensics.com/ от PAR Government Systems Corp.) для автоматизированного семантического анализа и поиска фейковых текстов, аудио, изображений, видео в реальном времени, с учетом атрибутов фейков (признаков злонамеренного умысла в содержании, манипулирования, характерных источников, способов/алгоритмов создания и управления дезинформацией). В России АНО «Диалог регионы» в 2023 году запустил платформу мониторинга аудиовизуальных дипфейков «Зефир» с помощью «алгоритмической оценки и анализа с помощью искусственного интеллекта», эффективность которой, по заявлению разработчиков, порядка 80%; поставлена задача разработки единой российской платформы для выявления ИИ-фейков. Недостаток нейросетевой технологии распознания фейков – их ограниченная эффективность (детекторы ИИ-контента, обученные искать контент конкретных ИИ-генераторов и определенные меркеры фейков, плохо справляются с другими типами фейков[25]). Например, проверка способности платформы известной

[25] Например, сервис hivemoderation.com/ai-generated-content-detection неплохо детектирует изображения, сгенерированные нейросетями, сервисы contentatscale.ai/ai-image-detector,

системы аутентификации лиц Facial Liveness Verification (FLV) отличить фейк от реального лица показала ее малоэффективность, причем система хуже людей распознавала очевидно нереальные видеофейки – и хуже всего детектировались подделки женских и цветных лиц (Li et al., 2022). Хотя нейросетвые модели способны во многих случаях определять подделку под реальное изображение, ИИ-инструменты имеют немалый процент ложных срабатываний, детектирование всегда немного отстает от инструментов изготовления фейков, по мере развития технологий.

4. Эспертные методы борьбы с фейками. Лишь технические методы борьбы с фейк-контентом недостаточны, чтобы общество могло противостоять дезинформации. Целесообразен комплексный подход, включающий не только программно-технические и законодательно-регуляторные меры противодействия фейкам, но и социально-психологические (в т.ч. формирование личных компетенций в распознавании фейков, рационального и критического подхода к потреблению информации в сетевых медиа,). Важно создание условий прозрачности и честности, когда работа журналистов и других создателей и ретрансляторов медиа-контента, согласуется с профессиональной этикой и нормами законов. Повышение общего уровня медиакомпетентности населения (опыт, «насмотренность», знание способов и инструментов проверки достоверности информации) помогают распознавать дезинформацию тем, кто производит контент и ретранслирует его. Знание базовых принципов и методов фактчекинга полезно не только журналистам, медиа-специалистам, исследователям, но и студентам и учащимся, активно потребляющим информацию из цифровых медиа и иных источников, формирующих на основе этой информации убеждения и мировоззрение. Среди методов фактчекинга – ручной и инструментальный поиск первоисточников, проверка авторов и цитируемых экспертов; OSINT-инструменты (whois.domaintools.com, liveuamap.com, tineye.com и др.), позволяющие узнать данные о цифровых

huggingface.co/spaces/umm-maybe/AI-image-detector, huggingface.co/spaces/umm-maybe/AI-image-detector хуже различают реальные фото и сгенерированный контент.

ресурсах, размещающих контент. Полезны фактчекинговые сервисы разных стран (Mediakritika.by, Snopes.com, FactCheck.org, trooclick.com, lapsha.media, politi-fact.com, aiwiz.ru/ai_fact_checking и др.); инструменты обратного поиска (например, функции поиска по картинке браузеров). Есть программы и сервисы для проверки социальной информации об авторах и персонах (peoplefinders.com, pipl.com, Wink, Spokeo, Email Lookup и др.), сайты официальной информации об организациях, людях; доверенные источники и агрегаторы новостей, аналитической, научно-образовательной информации. При оценке источников значимы их локация, время/дата упоминания о проверяемом событии; активность, характеристики, репутация, занимаемая идейная позиция медиа-ресурса; квалификация и возможная мотивация источника / автора; правильность (и контекст употребления) цитат, данных, фамилий, должностей и т. п. Полезны такие навыки фактчекинга, как внимание к подаче и структуре сообщения, способность распознать тональность и признаки манипулятивности сообщения (фейк-новости зачастую апеллируют к эмоциям, имеют оттенок сенсационности, алогичности, ангажированности); умение проанализировать иллюстрации в сообщении (в фейк-сообщениях они обычно не имеют отношения к новости, не имеют автора или сгенерированы нейросетью). Для сгенерированных нейросетью изображений характерны некоторые признаки, позволяющие усомниться в их достоверности: неестественная регулярность объектов, многократное клонирование элементов (много пальцев, однотипность предметов, поз, лиц, ракурсов в одном изображении); чем больше объектов, тем ниже качество детализации; стилизованность (например, пасторальная или апокалиптическая картинность); неестественная гладкость или, наоборот, странные структуры поверхностей; чрезмерная яркость, четкость всех элементов изображения – или, наоборот расфокусированность деталей, которые должны быть в фокусе; неестественное освещение и положение теней, отражений (в зрачках нет отражения, или оно различается в разных глазах); постановочный характер сцен и поз (обусловленный тем, что зачастую обучающий набор GNN-моделей состоит из фотографий фотостоков и соцсетей, а на таких фото кадр

специально выстраивается). При анализе видео на предмет его фейковости специалисты обычно обращают внимание на движения головы, губ, глаз, частоту моргания; расхождение темпа речи и движения губ может быть свидетельством фейковости, как и странное направление взгляда, регулярность/нерегулярность морганий, неестественные («механические») микродвижения тела при дыхании и т. д. Хотя создатели нейросетей постоянно дообучают модели, чтобы добиться большой реалистичности, цифровой опыт, насмотренность, избирательность к потреблению контента, критичность к содержанию информации и определенным видам источников, как общая установка, помогают обнаруживать фальсификацию, не вовлекаться в обман и манипуляцию, лучше ориентироваться в потоке сомнительной информации. Потребителям информации, вероятно, придется в будущем руководствоваться принципом «нулевого доверия» в отношении любой информации из цифровых медиа.

Чтобы побудить собственников медиа-ресурсов ответственно фильтровать размещаемый контент и пресекать распространение фальсифицированной информации, целесообразно формализовать признаки фейка, определить измеримые характеристики вредоносного фейк-контента в медиа-ресурсах, методики расчета показателей достоверности и фейковости. Содержательные признаки фейков поддаются контент-анализу; ссылки, цитаты, документы, фото и видео в сообщении можно проверить автоматически; упоминаемые люди и события могут быть распознаны и проверены по сведениям из доверенных СМИ и факт-чекинговых ресурсов, тематических и справочных хранилищ данных. Поведенческими признаками фейка могли бы служить: виральность контента (высокая частота его цитирований определенными сайтами, ботами, аккаунтами, ранее замеченными за рекламной и пропагандистской деятельностью с использованием фейков); динамика посещаемости ресурса (резкие колебания посещаемости или активности посетителей, их IP-адреса, среднее время чтения, реакция на контент – дизлайки, жалобы); среднее число подписчиков и отписок в единицу времени (люди быстро обнаруживают, если ресурс злоупотребляет фейк-контентом, и перестают посещать его через несколько дней или недель

чтения). Одним из измерителей фейковости может служить рейтинг доверия ресурсу со стороны авторизованных читателей, экспертов. Автоматизированные инструменты (в т. ч. средства фактчекинга) могли бы индексировать web-страницы по уровню достоверности/фейковости, подобно роботам-краулерам браузеров или аналитических SSM-инструментов. Может быть создан рейтинг-реестр в пределах национального сегмента интернета, отслеживающий интегральный уровень достоверности сетевых медиа-ресурсов. Большинство людей не приемлет дезинформации и манипуляций, и такого рода инструмент, как представляется, может быть востребован.

Выводы

Одних лишь технических решений для противодействия распространению фейкам недостаточно. Причины доверия к фейкам и стремления их распространять лежат в социально-психологической сфере. Что побуждает людей проводить время за поглощением и репостингом определенного (в т.ч. недостоверного) контента в медиапространстве, – тревога, скука, отсутствие близкодействующих социальных контактов и недостаток реального общения, глубоких интересов и реального дела, внушаемость, одиночество, бессмысленность жизни? Тревожность или неопределенность окружающей обстановки могут вызывать желание «сёрфить» интернет в поисках подтверждения или опровержения тревожащей информации; люди в состоянии тревожности (или наоборот, расслабленности) – более легкая жертва кибермошенников, использующих фейк-контент и приемы социальной инженерии. Ответы на вопросы о причинах доверчивости к фейк-контенту, склонности поглощать и транслировать в своем окружении информацию, не проверяя ее, умножая ложь, – лежат в социально-психологической сфере. Как и в случае с эпидемиями, противодействовать «заражению» и распространению фейк-контента мог бы когнитивный «иммунитет» к нему (индивидуальный и социальный). Индивидуальный «иммунитет» к фейкам образуют: стремление самореализоваться в реальном мире и в продуктивной деятельности, а не в виртуальной информационной среде; развитый уровень критичности,

сознательности, избирательности при потреблении информации; определенный жизненный опыт или сложившиеся убеждения; знание признаков фейков; повышенный порог внушаемости; общая медиаграмотность и цифровая грамотность; знание основных приемов социальной инженерии, умение прерывать манипуляцию. Чем меньше ретрансляторов фейк-контента в коммуникационных средах, тем меньше негативные социальные последствия фейков. Сознательный настрой большинства коммуникаторов: не выступать в роли ретранслятора недостоверной информации, – способен снизить количество фейк-контента и его разрушительные социальные эффекты. Помочь усилению индивидуального «иммунитета» к фальсификации и повышению осознанности могут гуманитарные технологии[26]. В исследование механизмов фальсификации (в т.ч. методологии противодействия фальсификации) должны активнее вовлекаться не только информационные и технические, но и когнитивные науки.

Фейки и методы социальной инженерии, подменяющие собой правдивые новости и честную коммуникацию, подрывают доверие к социальным институтам, медиаканалам, к социальному взаимодействию, способны разрушать согласие, разделяя и атомизириуя общество. Важен комплексный подход к противодействию фейк-контенту в цифровом пространстве. Толерантности к фейкам как медиа-явлению способствует в обществе «релятивистская» идеология (результатом которой стала идеология постправды, – в своем крайнем проявлении ценностно уравнивающая истину и ее подмену, фактоид). «Релятивизму» в трактовке событий в работе журналистов, исследователей должен противостоять принцип истины, как установка предельно честного отражения фактов и событий, важная и профессиональной журналистики, и для цифровых масс-медиа, если они хотят оставаться средством человеческой коммуникации, заслуживающими внимания, доверия и влияния.

[26] Например, методы игрификации (Roozenbeek, Linden, 2019, 65), позволяющие пользователю примерить на себя роли создателя фейка и фактчекера, чтобы понять «изнутри» процессы дезинформации и способы уклонения от нее.

Список литературы

1. CHugrov S.V. (2017). Post-truth: transformaciya politicheskoj real'nosti ili samorazrushenie liberal'noj demokratii? Polis. Politicheskie issledovaniya, (2), 42–59. https://doi.org/10.17976/jpps/2017.02.04. EDN YJEEWL

2. Ipsos: Survey on the impact of online disinformation and hate speech. Ipsos, 9 [Electronic resource]. – Access mode: https://www.unesco.org/sites/default/files/medias/fichiers/2023/11/unesco_ipsos_survey.pdf

3. Fitch A. (2019) Readers Beware: AI Has Learned to Create Fake News Stories [Electronic resource]. – Access mode: https://www.wsj.com/articles/readers-beware-ai-has-learned-to-create-fake-news-stories-11571018640].

4. Gowal S., Kohli P. (2023). Identifying AI-generated images with SynthID [Electronic resource]. – Access mode: https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid

5. Habermas J. (2022). Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik. Suhrkamp Verlag. Berlin.

6. Jung T., Kim S., Kim K. (2020). DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. IEEE Access. https://doi.org/10.1109/ACCESS.2020.2988660

7. Kirchner J.H., Ahmad L., Aaronson S., Leike J. (2023). New AI classifier for indicating AI-written text [Electronic resource]. – Access mode: https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text

8. Li C., Wang L., Ji S, Zhang X., Xi Z., Guo S., Wang T. (2022). Seeing is Living? Re-thinking the Security of Facial Liveness Verification in the Deepfake Era. arXiv:2202.10673v1. https://doi.org/10.48550/arXiv.2202.10673

9. Lukeš D., Laurent X., Pritchard J., Sharpe R., Walker C. (2023) Beyond ChatGPT: The state of generative AI in academic practice for autumn 2023. University of Oxford, 3–7 [Electronic resource]. – Access mode: https://www.ctl.ox.ac.uk/sites/default/files/ctl/documents/media/beyond_chatgpt_-_state_of_ai_for_autumn_2023_correct.pdf

10. Magesh V., Surani F., Dahl M., Suzgun M., Manning C.D., Ho D.E. (2024). Hallucina-tion-Free? Assessing the Reliability of Leading AI Legal Research Tools. https://doi.org/10.48550/arXiv.2405.20362

11. Marra F., Gragnaniello D., Verdoliva L. (2018). Detection of GAN-Generated Fake Im-ages over Social Networks. 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). https://doi.org/10.1109/MIPR.2018.00084

12. Newman N. (2024). Overview and key findings of the 2024 Digital News Report. Reuters Institute for the Study of Journalism [Electronic resource]. – Access mode: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024/dnr-executive-summary

13. Pluckrose H., Lindsay J.A., Boghossian P. (2018). Academic Grievance Studies and the Corruption of Scholarship [Electronic resource]. – Access mode: https://web.archive.org/web/20181013153454/https://areomagazine.com/2018/10/02/academic-grievance-studies-and-the-corruption-of-scholarship

14. Roozenbeek J., Linden S. (2019). Fake news game confers psychological resistance against online misinformation. Palgrave Communications, (5). https://doi.org/10.1057/s41599-0190279-9

15. Taylor R., Kardas M., Cucurull G., Scialom T., Hartshorn A., Saravia E, Poulton, A., Kerkez V., Stojnic R. (2022). Galactica: A Large Language Model for Science [Electronic resource]. – Access mode: https://arxiv.org/abs/2211.09085; https://doi.org/10.48550/arXiv.2211.09085

16. Vosoughi S., Roy D., Aral S. (2018). The Spread of True and False News Online. Science, (359), 1146–1151. https://doi.org/10.1126/science.aap9559

17. Weber-Wulff D., Anohina-Naumeca A., Bjelobaba S., Foltýnek T., Guerrero-Dib J., Po-poola O., Šigut P., Waddington L. (2023). Testing of Detection Tools for AI-Generated Text. International Journal for Educational Integrity, (19). https://doi.org/10.1007/s40979-023-00146-z. EDN VCYINN

18. Wei J., Tay Y., Bommasani R., Raffel C., Zoph B., Borgeaud S., Yogatama D., Bosma M., Zhou D., Metzler D., Chi E.H., Hashimoto T., Vinyals O., Liang P., Dean J.,

Fedus W. (2022). Emergent Abilities of Large Language Models. Published in Transactions on Machine Learning Research. https://doi.org/10.48550/arXiv.2206.07682

19. Фейк-ньюс – и как с ними бороться? // Сайт ВЦИОМ [Электронный ресурс]. – Режим доступа: https://wciom.ru/analytical-reviews/analiticheskii-obzor/feik-njus-i-kak-s-nimi-borotsja (дата обращения: 22.11.2023).

20. Грачев М.Н. Концепт «разрушения правды» в условиях цифрового общества (аналитический обзор) / М.Н. Грачев, Р.В. Евстифеев // Контуры глобальных трансформаций: политика, экономика, право. – 2020. – №13 (2). – С. 229–248. https://doi.org/10.23932/2542-0240-2020-13-2-12. – EDN CKTGBY

21. Ершов Ю.М. Феномен фейка в контексте коммуникационных практик / Ю.М. Ершов // Вестник Томского государственного университета. Филология. – 2018. – №52. – С. 245–256. https://doi.org/10.17223/19986645/52/15. – EDN XPICLR

22. Кириллова Н.Б. Цифровая культура как новая социально-антропологическая реальность и проблемы идентичности / Н.Б. Кириллова // Современная наука: актуальные проблемы теории и практики. Серия: Познание. – 2023. – №7. – С. 13–18. https://doi.org/10.37882/2500-3682.2023.07.04. – EDN MEOULJ

23. Тумбинская М.В. Идентификация фейк-новостей с помощью веб-ресурса на основе нейронных сетей / М.В. Тумбинская, Р.А. Галиев // Программные продукты и системы. – 2023. – №4. – С. 590–599. https://doi.org/10.15827/0236-235X.144.590-599. – EDN AEJSLD

24. Указ Президента РФ от 10.10.2019 г. №490 «О развитии искусственного интеллекта в Российской Федерации» // СПС Гарант [Электронный ресурс]. – Режим доступа: https://base.garant.ru/72838946 (дата обращения: 17.10.2024)

25. Ушкин С.Г. Не только социальные сети: каналы распространения фейковых новостей в представлениях населения / С.Г. Ушкин // Галактика медиа. Журнал медиа исследований. Исследовательский электронный журнал. – 2024. – №6 (2). – С. 162–176. https://doi.org/10.46539/gmd.v6i2

26. Чугров С.В. Post-truth: трансформация политической реальности или саморазрушение либеральной демократии? / С.В. Чугров // Полис. Политические исследования. – 2017. – №2. – С. 42–59. https://doi.org/10.17976/jpps/2017.02.04. – EDN YJEEWL

Миронова Наталия Геннадьевна – канд. филос. наук, доцент Института информатики, математики и робототехники ФГБОУ ВО «Уфимский университет науки и технологий», Уфа, Россия.