

Искусственный интеллект (ИИ) и распространение ложной информации

Изабель Брюер

Дезинформация и ложная информация широко распространены. В период с 2013 по 2019 год в социальных сетях было проведено 96 кампаний, призванных повлиять на результаты политических процессов за рубежом и внутри этих стран путем распространения ложной информации.¹ 20 февраля 2020 года Всемирная организация здравоохранения (ВОЗ) объявила о глобальной эпидемии, определяемой как “цунами информации — частично достоверной, частично нет, — которая распространяется параллельно с эпидемией”.² Эта инфодемия усилилась из-за притока дипфейков (от английского *deep fake*) — реалистичных изображений, видео- или аудио-контента, созданных с использованием “глубокого” машинного обучения и технологий ИИ. Хотя дипфейки имеют сходства с фоторедакторами и онлайн-фильтрами, они принципиально отличаются друг от друга, поскольку используют алгоритмы глубокого обучения для создания гиперреалистичных синтетических медиа, практически неотличимых от аутентичного контента. Хотя дипфейки могут быть безвредны, если их использовать исключительно в развлекательных целях, таких как создание версии фильма “Назад в будущее” с участием актерского состава из фильма “Человек-паук” 2017 года, они также могут использоваться для распространения ложной информации, разжигания ненависти и поляризации общества.

Появление дипфейков совпало с расширением использования социальных сетей и доступа к Интернету, что усилило распространение мисинформации и дезинформации (ложной информации без намерения причинить вред и ложной информации, направленной на причинение вреда, соответственно). Число пользователей Интернета во всем мире выросло с 416 миллионов в 2000 году до 4,7 миллиарда за каких-то 20 лет.³ С ростом числа дипфейков за последнее время ложная информация и дезинформации все чаще становятся неотличимыми от реальной информации.

На личном уровне технологии ИИ могут причинять вред, поскольку дипфейки позволяют даже неквалифицированным людям создавать фальшивый порнографический контент, который может быть использован для унижения, преследования и шантажа. Компания *Diptrace*, занимающаяся разработкой искусственного интеллекта, сообщила, что 96% дипфейков, циркулирующих в сети, являются порнографическими, при этом лица знаменитых женщин накладываются на тела порноактрис. Профессор права Бостонского

¹ Martin, Diego A, et al. “Introducing the Online Political Influence Efforts Dataset.” *Journal of Peace Research*, 10 Nov. 2022, p. 002234332210928, <https://doi.org/10.1177/00223433221092815>. Accessed 20 Jan. 2023.

² “Call for Action: Managing the Infodemic.” *Www.who.int*, 11 Dec. 2020, www.who.int/news/item/11-12-2020-call-for-action-managing-the-infodemic.

³ Murphy, Julia, et al. “Internet.” *Our World in Data*, 2023, ourworldindata.org/internet.

университета США (Boston University) Даниэль Ситрон (Danielle Citron) утверждает, что “технология дипфейков используется в качестве оружия против женщин”.⁴

Как отмечает ВОЗ, распространение дипфейков в сочетании с бурным ростом социальных сетей и цифровизацией сделало распространение фальшивых новостей во время глобальной пандемии особенно опасным. В интернете активно распространялись ложные сведения о том, что алкоголь, экстремальная жара или холод могут защитить от вируса, несмотря на отсутствие подтверждающих эти утверждения научных доказательств.⁵ В начале пандемии Центры по контролю и профилактике заболеваний (CDC) сообщили о результатах опроса, согласно которому 39% респондентов подвергали себя опасным воздействиям, связанных с использованием бытовых дезинфицирующих средств, включая “мытьё пищевых продуктов отбеливателем, нанесение бытовых чистящих средств непосредственно на кожу и намеренное вдыхание или употребление дезинфицирующих средств для предотвращения заражения COVID-19”.⁶

Политическое использование дипфейков — например, видео, в котором Барак Обама называет Дональда Трампа “полным идиотом”⁷, или ролик, где речь политика-демократа Нэнси Пелоси, была слегка замедлена, чтобы ее речь казалась невнятной, — иллюстрирует более тревожные последствия искусственного интеллекта и дезинформации.⁸ Исследование, в котором сравнивалось распространение фейковых и правдивых новостей в Интернете, показало, что дезинформация в Twitter распространяется на 70% быстрее, чем реальная информация, при этом ложные политические новости являются наиболее вирусными.⁹

Подделки с использованием искусственного интеллекта также представляют серьезную угрозу международной безопасности. В марте 2022 года, вскоре после российского вторжения в Украину, российские пропагандисты сфабриковали видео, в котором Президент Украины Владимир Зеленский призывает своих военных сдаться. Этот фейк быстро завоевал популярность в социальных сетях и новостях, вызвав замешательство и обострив напряженность в разгар кризиса.¹⁰ Офис Президента Зеленского быстро опроверг подлинность видео, но этот случай использования дипфейков

⁴ Sample, Ian. “What Are Deepfakes – and How Can You Spot Them?” *The Guardian*, 13 Jan. 2020, www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them.

⁵ Group, Internet. “Deepfakes and the Healthcare Industry during COVID-19.” *GlobalSign*, 2 Apr. 2024, www.globalsign.com/en/blog/in/how-deepfakes-could-further-burden-healthcare-industry-during-covid-19. Accessed 2 Jan. 2025.

⁶ Nelson, Taylor, et al. “The Danger of Misinformation in the COVID-19 Crisis.” *Missouri Medicine*, vol. 117, no. 6, Nov. 2020, p. 510, [pmc.ncbi.nlm.nih.gov/articles/PMC7721433/](https://pubmed.ncbi.nlm.nih.gov/articles/PMC7721433/).

⁷ Sample, Ian. “What Are Deepfakes – and How Can You Spot Them?” *The Guardian*, 13 Jan. 2020, www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them.

⁸ Mervosh, Sarah. “Distorted Videos of Nancy Pelosi Spread on Facebook and Twitter, Helped by Trump.” *The New York Times*, 24 May 2019, www.nytimes.com/2019/05/24/us/politics/pelosi-doctored-video.html?searchResultPosition=7.

⁹ Vosoughi, Soroush, et al. *FALSE NEWS IS BIG NEWS*. 2017.

¹⁰ Buffett Brief. *The Rise of Artificial Intelligence and Deepfakes*. July 2023.

в вооруженных конфликтах показывает, что подобные угрозы международной безопасности будут возникать и в будущем, если не принять меры для предотвращения таких кризисов. Несмотря на катастрофические последствия дипфейков, исследователи из Северо-Западного университета США (Northwestern University) предполагают, что при наличии строгого кодекса поведения для правительств дипфейки могут быть использованы для противодействия и дестабилизации террористических группировок или других вредоносных организаций.

Появление дипфейков и осознание их пагубного воздействия на общество привело к росту числа исследований, предлагающих способы защиты от этих угроз. Разработанная DeepTrust Alliance “Система противодействия дезинформации” описывает возможные решения проблемы дезинформационной инфодемии и других последующих кризисов, вызванных искусственным интеллектом и дипфейками. Рамочная программа предлагает четыре основные стратегии: технологическое развитие (например, технологии распознавания), государственное регулирование (например, законодательство), частные нормативные акты (например, условия предоставления услуг) и общественное просвещение.¹¹ Такие компании, как Microsoft и Facebook, внедрили инструменты искусственного интеллекта для выявления дипфейков с повышенной точностью — некоторые инструменты достигли 80-90% эффективности.¹² Закон "О запрете вредоносных дипфейков" 2018 года, принятый 115-м Конгрессом США, предусматривает уголовную ответственность для лиц, создающих или распространяющих дипфейки с целью причинения вреда, например, путем распространения ложных политических материалов или фальшивого порнографического контента.¹³

¹¹ Yamaoka-Enkerlin, Anna. *DISRUPTING DISINFORMATION: DEEPFAKES and the LAW*. 8 Oct. 2020.

¹² Burt, Tom. “New Steps to Combat Disinformation.” *Microsoft on the Issues*, 1 Sept. 2020, blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/.

¹³ Sasse, Ben. *Malicious Deep Fake Prohibition Act of 2018*. 21 Dec. 2018.