ФЕЙКИ И ДИПФЕЙКИ В ИНТЕРНЕТЕ: БОРЬБА ПО ПРИНЦИПУ АЙКИДО

О.Л. Фиговский
Доктор технических наук, Академик EAS, РИА и РААСН, Президент Ассоциации изобретателей Израиля, Глава Департамента науки, технологий и образования Альянса Народов Мира, Израиль, figovsky@gmail.com

О.Г. Пенский
Доктор технических наук, профессор Пермского государственного национального исследовательского университета, Россия, ogpensky@mail.ru

В статье приводятся критические численные значения количества размещаемых в Интернете фейков, способных превратить интернет в абсурд; описываются дипфейки; предлагается способ борьбы с фейками и дипфейками, основанный на математической теории пресыщения эмоционального воспитания и гипотезе психолога Д.Н. Узнадзе и использующий принцип борьбы айкидо.

Ключевые слова: фейки, дипфейки, интернет, информационная безопасность, психология, воспитание, пресыщение воспитания.

FAKES AND DIPFAKES ON THE INTERNET: FIGHTING ON THE PRINCIPLE OF AIKIDO

O. Figovsky
Doctor of Technical Sciences, Academician of EAS, RIA and RAASN, President of the Israel Inventors Association, Head of the Department of Science, Technology and Education of the Alliance of the Nations of the World, Israel, figovsky@gmail.com

O. Pensky
Doctor of Technical Sciences, Professor, Perm State University, Russia, ogpensky@mail.ru

The article provides critical numerical values of the number of fakes posted on the Internet that can turn the Internet into absurdity; deepfakes are described; a method of combating fakes and deepfakes is proposed, based on the mathematical theory of satiety in upbringing and the hypothesis of psychologist D.N. Uznadze and using the principle of fighting aikido.

Key words: fakes, deepfakes, internet, information security, psychology, education, satiety of education.

Введение

В настоящее время глобальная сеть интернет стала очень популярной среди многих людей из всех социальных слоев. Но следует отметить то, что, на наш взгляд, эта сеть может в недалеком будущем потерять свою информационную значимость у людей и даже полностью превратиться в информационный абсурд. Причиной этого является огромный рост ложной информации, размещаемой и глобальной сети и называемой фейками и дипфейками.

Фейки

В статье «К вопросу о подлинности интернета: сколько в сети фейков и роботов» [https://www.hwp.ru//articles/K_voprosu_o_podlinnosti_interneta_skolko_v_seti_feykov_i_robotov_150557/] сказано следующее: Уже «…сегодня весь внешний оффлайновый мир с его новостями и скучными заголовками, воспринимается уже как иллюзия, жалкая пародия на настоящий мир онлайн-сервисов, таких как Youtube, Яндекс Дзен, Телеграм-каналы или Instagram. Мы верим блоггерам больше, чем профессиональным журналистам, мы вообще не доверяем ничему, что не можем проверить из «независимых источников», и самое интересное - если в комментариях мы встретим хотя бы 5 человек, чья точка зрения, совпадающих с нашей, то нас уже не переубедить… Но более 60% интернета - это фейк - обман, подделка, дорогая или дешёвая фальшивка, бот-ферма, кликбейт и т.д.». Количество фейков в интернете выросло в 10 раз в период пандемии [https://tass.ru/obschestvo/8673695]. Национальный центр помощи детям и Лига безопасного интернета РФ выявили более 33 тысяч фейков в сети с марта 2020 года. В 2021 году эксперты предсказывают двукратный рост фейков в интернете. Об этом РИА Новости сообщила член Общественной палаты России, директор центра Екатерина Мизулина.

Е. Мизулина сказала также следующее: «За 9 месяцев работы по мониторингу фейковой информации волонтерами Национального центра помощи детям и Лиги безопасного интернета обнаружено 33 тысячи недостоверных сообщений, распространенных под видом достоверных. Из них 17607 фейковых сообщений о коронавирусе, 8407 — об общероссийском голосовании по поправкам в Конституцию России, 3922 — о вакцине от коронавируса, 1781 — о едином дне голосования, 1312 — о введении локдауна в России на новогодние праздники».

Согласно данным мониторинга, подобные сообщения чаще всего распространялись в WhatsApp, YouTube, Viber, TikTok, Telegram, Instagram и «Одноклассниках». «По данным опроса, проведенного Лигой безопасного интернета, в декабре 2020 года более 82 процентов граждан сталкивались с недостоверными публикациями. При этом 71 процент заявили о том, что поверили полученной недостоверной информации в соцсетях и на

видеохостингах, 88 процентов говорят о том, что не стали сомневаться в информации, которую им переслали близкие и знакомые в мессенджерах», — сообщила директор центра [https://lenta.ru/news/2020/12/22/fake/].

В монографии [Пенский О.Г., Шарапов Ю.А., Ощепкова Н.В. Математические модели роботов с неабсолютной памятью и приложения моделей. Пермь: изд-во ПГНИУ. 2018. 365 с.] предложена математическая теория формирования общественного сознания с помощью средств массовой информации, в том числе информации ресурсов сети интернет. В этой книге математически строго доказано, что при непрерывном воздействии стимулами на аудиторию воспитание аудитории, обусловленное возникающими у аудитории положительными эмоциями, стремится к асимптоте, говоря математическим языком, «сходится». При приближении воспитания к асимптоте согласно гипотезе директора Института Психологии Академии наук Грузинской ССР Д.Н. Узнадзе воспитание может сменить знак, т.е., говоря математическим языком, его положительное значение начнет принимать только отрицательные численные значения. Основываясь на этой математике, можно заключить, что массовое распространение фейков в интернете, может привести к отрицательному отношению пользователей глобальной сети, ранее положительно относящихся к этой сети, то есть, интернет потеряет свою популярность у пользователей, как источник информации, что повлечет, например, неизбежное разорение владельцев некоторых интернет-порталов.

Если предположить, что количество фейков, размещаемых в сети интернет, пропорционально величине эмоционального воспитания аудитории, то согласно данным Е. Мизулиной, приведенным выше, можно вычислить критическое значение количества фейков в интернете, которое приведет к отрицательному эмоциональному восприятию информации, размещаемой в интернете, а, значит, и отрицательному отношению ко всей глобальной сети. Это критическое значение для России равно 140 000 фейкам, размещаемым за год.

Дипфейки

Обычно под фейками понимают ложную текстовую информацию.

Но относительно недавно, в основном, начиная с 2018 года, активизировалось размещение в интернете, так называемых, дипфейков. Отметим то, что в настоящее время в интернете уже существует множество сайтов, предлагающих услуги как в обучении созданию фейков и дипфейков, так и в их производстве.

Приведем общепринятое определение дипфейка.

Deepfake — конкатенация слов «глубинное обучение» (англ. deep learning) и «подделка» (англ. fake), методика синтеза изображения и звука, основанная на искусственном интеллекте (ИИ).

Исследователи из лаборатории SAND Lab при Чикагском университете утверждают [https://habr.com/ru/post/586794/], что доступные широкой общественности программы клонирования голоса развиваются угрожающе быстро. В частности, создаваемые при помощи подобных технологий

голосовые дипфейки могут сбить с толку как людей, так и смарт-устройства с голосовым управлением. Но не только голосовые, а также и другие дипфейки представляют угрозу. Дипфейки уже используются в рекламе, моде, журналистике и обучении. Однако больше 90% подделок созданы, чтобы навредить репутации, — например с помощью порнографических роликов. Общие денежные потери бизнеса из-за дипфейков приблизились к отметке $250 млн в 2020 году. Ниже приведем выдержки из статьи [https://habr.com/ru/post/586794/].

Команда экспертов из лаборатории SAND Lab (Security, Algorithms, Networking and Data Lab) протестировала доступные на платформе Github программы клонирования голоса, чтобы узнать, смогут ли они прорваться сквозь защиту системы безопасности распознавания голоса умных колонок Alexa, WeChat и Azure. Среди множества программ внимание учёных привлекла технология SV2TTS, создатели которой называют её «инструментом для клонирования голоса в режиме реального времени». По словам разработчиков, программа SV2TTS может синтезировать полноценные голосовые дипфейки, основываясь всего на 5 секундах записи оригинального голоса. Имитация голоса в исполнении SV2TTS смогла обойти защиту Azure от Microsoft в 30% случаев, а колонки Alexa и WeChat поддавались обману и того чаще – в 63%. Не менее ошеломляющие результаты показал эксперимент с 200 волонтёрами: примерно в половине случаев люди не смогли отличить голосовые дипфейки от реальных голосов. Кроме того, исследователи пришли к выводу, что по какой-то причине синтезаторам речи гораздо лучше удаётся имитировать женские голоса, а также речь людей, для которых английский язык не является родным. По словам исследователей, современные механизмы защиты против синтезированной речи развиваются медленнее технологий имитации голоса. Не в тех руках подобные программы рискуют стать инструментом воплощения преступного замысла. При этом мишенью атаки могут стать как реальные люди, так и смарт-устройства. К примеру, колонка WeChat использует распознавание голоса пользователя для обеспечения доступа к платным функциям, например, для проведения транзакций в приложениях сторонних разработчиков вроде Uber или New Scientist.

Как было отмечено выше, зачастую цель с фото или видео дипфейков — это порнографический шантаж. Жертвами регулярно становятся знаменитости. Например, в числе жертв злоумышленников — Натали Портман, Тейлор Свифт, Галь Гадот и другие. Но подделывают внешность и голос не только известных людей.

Другой интересный пример: видео с экс-премьер-министром Бельгии Софи Вильмес, которое опубликовала организация Extinction Rebellion Belgium. Этот дипфейк — модификация предыдущего обращения к нации по поводу пандемии. В вымышленной речи говорится, что последние глобальные эпидемии напрямую связаны с «эксплуатацией и разрушением людьми природной среды». Для воспроизведения голоса и манеры речи использовался ИИ.

Раньше можно было спокойно отличить дипфейк по различным признакам:

- Неестественная интонация речи;
- Роботизированный тон;
- Неестественное моргание или движение человека на видео;
- Движения губ не синхронизированы с речью;
- Низкое качество звука или видео;

Но сейчас — это сложно. Противодействие дипфейкам — игра в кошки-мышки, так как мошенники тоже совершенствуют технологию. Например, в 2018 году выяснилось, что люди в дипфейках не моргают или делают это странно. Особенность сразу же учли в усовершенствованных моделях.

Чтобы стимулировать создание технологий для обнаружения дипфейков, Facebook и Microsoft проводят Deepfake Detection Challenge. В 2020 году в нем приняли участие более 2 тыс. человек. Разработчикам удалось добиться точности распознавания более 82% на стандартном тестовом датасете, но на усложненном (с отвлекающими компонентами вроде надписей) она упала чуть более чем до 65%.

Программное обеспечение для обнаружения дипфейков можно обмануть, слегка видоизменяя входные данные, так что исследователи продолжают вести работу в этой области. В июне 2021 года ученые из Facebook и Университета штата Мичиган объявили о новой разработке. Обычно детекторы определяют, какая из известных моделей ИИ сгенерировала дипфейк. Новое решение лучше подойдет для практического применения: оно может распознавать подделки, созданные с помощью методов, с которыми алгоритм не сталкивался при обучении.

Еще один способ: поиск цифровых артефактов. У людей в дипфейках могут не совпадать цвета левого и правого глаза, расстояние от центра глаза до края радужной оболочки, отражение в глазах. Встречаются плохо прорисованные зубы и нереалистично темные границы носа и лица. Но в современных подделках увидеть такие артефакты может только машина. Так что самое важное в борьбе с дипфейками — быть настороже и обращать внимание на фото, видео и аудио, которые кажутся подозрительными.

Доступность датасетов и предобученных нейросетевых моделей, снижение стоимости вычислений и соревнование между создателями и детекторами дипфейков подгоняет рынок. Deepfake-инструменты коммодифицируются: в сети свободно распространяются программы и учебные материалы для создания подделок. Существуют простые приложения для смартфонов, которые вообще не требуют технических навыков.

Повышенный спрос на дипфейки привел к созданию компаний, которые предлагают их как продукт или услугу. Это, например, Synthesia и Rephrase.ai. В Ernst & Young дипфейки, сделанные с помощью Synthesia, уже начали использовать в клиентских презентациях и в переписке. Стартап Sonantic специализируется на озвучке видеоигр и предлагает платформу no-code для генерации голосовых клонов.

Вместе с рынком коммерческого применения дипфейков будет расти и число мошеннических операций. В сети уже есть маркетплейсы, где публикуют запросы на подделки, например порноролики с актрисами. А некоторые алгоритмы могут сгенерировать deepfake-видео на основе одного изображения или воссоздать голос человека, используя аудио длиной в несколько секунд. Из-за этого жертвой аферистов может стать практически любой пользователь интернета.

С дипфейками борются не только с помощью технологий, но и на законодательном уровне. В США и Китае появляются законы, регулирующие их использование, а в России борьбу с ними в июле 2021 года включили в одну из дорожных карт «Цифровой экономики». Регулирование в этой области будет только усиливаться.

Мы из-за отсутствия данных о ежегодном количестве размещенных дипфейков в интернете не можем привести вычисленное критическое число дипфейков, способное превратить видео-интернет в абсурд, но, основываясь на общих закономерностях математического эмоционального воспитания, которое имеет пресыщение, можем предложить следующий способ психологического избавления аудитории от влияния дипфейков и одновременно спасения интернета как правдивого источника аудио-и-видео информации.

Для этого достаточно создать официальный портал, куда нужно размещать все дипфейки, размещаемые в интернете. По началу любители подделок будут активно просматривать и прослушивать дипфейки, размещенные на портале, но потом из-за воспитательного пресыщения дипфейками им попросту просмотр дипфейков надоест (математическая формула параметра «надело» приведена, например, в монографии [Пенский О.Г., Шарапов Ю.А., Ощепкова Н.В. Математические модели роботов с неабсолютной памятью и приложения моделей. Пермь: изд-во ПГНИУ. 2018. 365 с.]. В результате пресыщения дипфейками «любознательной» аудитории положительное восприятие дипфейков согласно гипотезе Д.Н. Узнадзе сменится на отрицательное восприятие. Производители дипфейков потеряют большую часть рынка сбыта, а поэтому размещение дипфейков в сети интернет значительно уменьшится.

Кстати, для сокращения производства чисто фейковой информации можно создать такой же официальный портал фейков.

Отметим, что по своей сути идея функционирования порталов фейков и дипфейков та же, что и принципы борьбы айкидо, где направление удара противника используется для того, чтобы победить этого противника.

Заключение

Таким образом, в статье впервые предложен способ борьбы с фейками и дипфейками в интернете, основанный на гипотезе грузинского психолога Д.Н. Узнадзе и формализованной математической теорией эмоциональных роботов. Конечно, для внедрения в жизнь интернета специализированных порталов фейков и дипфейков (при условии, что их содержимое не противоречит законодательствам государств) необходимы дополнительные натурные испытания на репрезентативных группах людей. Но авторы настоящей статьи

думают, что эти испытания подтвердят правильность изложенных выше предложений. Отметим то, что интернет-ресурсы, на которые ссылаются в статье авторы, фейками не являются, так как эти ресурсы есть источники, прошедшие дополнительную проверку.