**Что такое дипфейки и как их распознать?**

Поддельные видео, созданные искусственным интеллектом, становятся все более распространенными и убедительными. Вот почему нам следует волноваться.

**Что такое дипфейк?**

Вы видели, как Барак Обама называл Дональда Трампа «полным дураком», или Марк Цукерберг хвастался тем, что он «полностью контролирует миллиарды людей украденными данными», или были свидетелями трогательных извинений Джона Сноу за мрачный финал Игры престолов? Ответьте да, и вы видели дипфейк.

Ответ 21 века на фотошоппинг: дипфейки используют форму искусственного интеллекта, для создания изображений фейковых событий, отсюда и название дипфейка. Хотите вложить новые слова в уста политика, сняться в любимом фильме  или потанцевать как профессионал? Тогда пора делать дипфейк.

**Для чего они?**

Многие из них являются порнографическими. В сентябре 2019 года компания Deeptrace, занимающаяся искусственным интеллектом, обнаружила в сети 15000 дипфейков, что почти вдвое больше, чем за девять месяцев. Ошеломляющие 96% были порнографические и в 99% случаев фигурируют лица знаменитостей женского пола якобы работающие порнозвездами.

Новые технологии позволяют неквалифицированным людям делать deepfake из горстки фотографий. Поддельное видео, скорее всего, распространили как порноместь. Но технология Deepfake используется не только против женщин.

 **Это только о видео?**

Нет. Технология Deepfake позволяет создавать убедительные, но полностью вымышленные фотографии с нуля. Несуществующая журналистка Bloomberg Мэйзи Кинсли, у которой был профиль в LinkedIn и Twitter, вероятно, была дипфейком.

Другая подделка LinkedIn - Кэти Джонс, утверждала, что работает в Центре стратегических и международных исследований, но она считается дипфейком, созданным для иностранной шпионской операции.

Звук тоже может быть deepfake, чтобы создать «голосовые клоны» общественных и/или политических деятелей. В марте прошлого года руководитель британской дочерней немецкой энергетической компании внес почти 200 000 фунтов стерлингов на счет в венгерском банке после того, как ему позвонил мошенник, имитирующий голос немецкого генерального директора. Страховые компании считают, что голос был подделкой, но доказательства не выяснены. Сообщается, что аналогичные мошенничества использовали записанные голосовые в сообщениях WhatsApp.

 Сравнение оригинального и deepfake видео исполнительного директора Facebook Марка Цукерберга. Фотография: The Washington Post через Getty Images.

**Как они сделаны?**

Университетские исследователи и студии спецэффектов давно раздвинули границы возможностей обработки видео и изображений. Но deepfake сами родились в 2017 году, когда один пользователь под именем Reddit выложил порно клипы на сайте. В видео поменяли лица и вставили лица знаменитостей - Гадот, Taylor Swift, Скарлетт Йоханссон и другие - на порно исполнителей.

***Как это достигается?***

Чтобы сделать видео со сменой лица, нужно выполнить несколько шагов. Сначала запускается тысячи снимков лиц двух людей с помощью алгоритма искусственного интеллекта, называемого кодировщиком.

Кодировщик находит и изучает сходства между двумя лицами и сводит их к их общим чертам, сжимая изображения в процессе.

Второй алгоритм, называемый декодером, обучается восстанавливать лица из сжатых изображений. Поскольку лица разные, один декодер для восстановления лица первого человека, а другой декодер - для восстановления лица второго человека. Чтобы выполнить замену лица, просто загружается закодированные изображения в «неправильный» декодер.

Например, сжатое изображение лица человека A загружается в декодер, обученный на человеке B. Затем декодер восстанавливает лицо человека B с выражениями и ориентацией лица A для получения убедительного видео.

Другой способ создания дипфейков - это так называемая генеративная состязательная сеть, или Gan. Ган противопоставляет два алгоритма искусственного интеллекта.

Первый алгоритм, известный как генератор, обрабатывает случайный кадр и превращает его в изображение. Затем это синтетическое изображение добавляется к потоку реальных изображений - скажем, знаменитостей, которые вводятся во второй алгоритм, известный как дискриминатор. Сначала синтетические изображения не будут похожи на лица. Но повторяйте процесс бесчисленное количество раз с обратной связью о производительности, и дискриминатор и генератор улучшатся. При наличии достаточного количества циклов и обратной связи генератор начнет создавать совершенно реалистичные лица совершенно несуществующих знаменитостей.

**Кто делает дипфейки?**

Каждый из академических и промышленных исследователей самодеятельных энтузиастов, визуальных эффектов студии и порно производителей. Правительства также могут баловаться этой технологией в рамках своих онлайн-стратегий, чтобы дискредитировать и разрушить экстремистские группы или, например, установить контакты с целевыми лицами.

**Какие технологии для этого нужны?**

Сделать хороший дипфейк на стандартном компьютере сложно. Большинство из них создается на высокопроизводительных настольных компьютерах с мощными видеокартами или, что еще лучше, с вычислительной мощностью в облаке. Это сокращает время обработки с дней и недель до часов.

Но для того, чтобы подправить готовые видеоролики, чтобы уменьшить мерцание и другие визуальные дефекты, также требуется опыт. Тем не менее, сейчас доступно множество инструментов, которые помогут людям создавать дипфейки. Несколько компаний сделают их для вас и сделают всю обработку в облаке. Есть даже приложение для мобильных телефонов, например, Zao, которое позволяет пользователям добавлять свои лица в список героев телешоу и фильмов.

Китайское приложение для обмена лицами Zao вызвало опасения по поводу конфиденциальности.

**Как определить дипфейк?**

По мере совершенствования технологий становится все труднее распознавать дипфейк. В 2018 году американские обнаружили, что лица deepfake не мигают нормально.

В этом нет ничего удивительного: большинство изображений показывают людей с открытыми глазами, поэтому алгоритмы никогда не узнают о моргании. Сначала это казалось проблемой обнаружения. Но как только исследование было опубликовано, появились дипфейки с морганием. Такова природа игры: как только слабость обнаруживается, она устраняется.

Дипфейки плохого качества обнаружить легче. Синхронизация губ может быть плохой, а оттенок кожи неоднородным. По краям лиц может наблюдаться мерцание. А мелкие детали, такие как волосы, особенно сложно для дипфейков, особенно там, где видны пряди на челке. Плохо обработанные украшения и зубы также могут быть ориентиром, как и странные световые эффекты, такие как непоследовательное освещение и отражения на радужной оболочке.

Правительства, университеты и технологические компании финансируют исследования для выявления дипфейков.

Не так давно стартовал первый конкурс Deepfake Detection Challenge при поддержке Microsoft, Facebook и Amazon. В него войдут исследовательские группы со всего мира, которые будут бороться за превосходство в игре по обнаружению дипфейков.

Facebook уже запретили дипфейк- видео, которые могут ввести зрителей в заблуждение, чтобы они подумали, что кто-то «сказал слова, которых они на самом деле не говорили», в преддверии выборов в США 2020 года. Однако политика распространяется только на дезинформацию, созданную с использованием ИИ (искусственного интеллекта), что означает, что на платформе все еще разрешены «мелкие подделки».

**Разрушат ли дипфейки?**

Мы можем ожидать новых дипфейков, которые беспокоят, запугивают, унижают, подрывают и дестабилизируют. Но вызовут ли дипфейки серьезные международные инциденты? Здесь ситуация менее ясна. Дипфейк, когда мировой лидер нажимает на большую красную кнопку, не должен вызывать армагеддон. Дипфейк-спутниковые снимки войск, сосредоточенных на границе, также не вызовут особых проблем: у большинства стран есть собственные надежные системы визуализации безопасности.

Впрочем, поводов для озорства еще достаточно. В прошлом году акции Tesla рухнули, когда Илон Маск выкурил косяк на онлайн-шоу.

В декабре Дональд Трамп рано улетел домой со встречи НАТО, когда появились подлинные кадры, на которых другие мировые лидеры, по-видимому, издеваются над ним. Смогут ли правдоподобные дипфейки изменить цены на акции, повлиять на избирателей и спровоцировать религиозную напряженность? Почему нет?

**Подорвут ли они доверие?**

Более коварное влияние дипфейков, наряду с другими синтетическими СМИ и фейковыми новостями, заключается в создании общества с нулевым доверием, в котором люди не могут или больше не заботятся о том, чтобы отличить правду от лжи. А когда доверие подрывается, легче усомниться в конкретных событиях.

В прошлом году министр связи Камеруна назвал фальшивыми новостями видео, на котором, по мнению Amnesty International, показано, как солдаты Камеруна убивать мирных жителей.

Дональд Трамп, который признался, что хвастался, что эксперт женских гениталий в записанном разговоре, позже заявил, что запись ненастоящая.

Проблема может быть не столько в подделке реальности, сколько в том, что реальность становится правдоподобно опровергнутой.

По мере того, как технология становится все более доступной, дипфейки могут создавать проблемы для судов, особенно в битвах за опеку над детьми и спорах по трудовым вопросам, где фальшивые события могут быть представлены в качестве доказательства.

Они также представляют угрозу для личной безопасности: дипфейки могут имитировать биометрические данные и потенциально могут обмануть системы, которые полагаются на распознавание лица, голоса, глаз или походки.

Возможность мошенничества очевидна. Позвоните кому-нибудь неожиданно, и они вряд ли переведут деньги на неизвестный банковский счет. Но что, если ваша «мама» или «сестра» сделает видеозвонок в WhatsApp и задаст тот же запрос?

**Какое решение?**

Как ни странно, ответом может быть Искусственный Интеллект. Искусственный интеллект уже помогает обнаруживать фальшивые видео, но у многих существующих систем обнаружения есть серьезный недостаток - они лучше всего подходят для знаменитостей.

Многие специализированные фирмы сейчас работают над системами обнаружения, которые нацелены на обнаружение подделок всякий раз, когда они появляются.

Другая стратегия фокусируется на происхождении в СМИ. Цифровые водяные знаки не являются надежными, но онлайн-бухгалтерская система блокчейна может хранить защищенные от несанкционированного доступа записи видео, изображений и аудио, поэтому их происхождение и любые манипуляции всегда можно проверить.

**Всегда ли дипфейки злы?**

Нет не всегда. Многие из них занимательны, а некоторые полезны. Дипфейки с клонированием голоса могут восстанавливать голоса людей, потерявших их из-за болезни.

Видео Deepfake могут оживить галереи и музеи. Во Флориде в музее Дали есть дипфейк художника-сюрреалиста, который представляет свое искусство и делает селфи с посетителями.

Для индустрии развлечений, технология может быть использована для улучшения перезаписи на иноязычных фильмах или воскрешать мертвых актеров. Например, покойный Джеймс Дин должен сыграть главную роль в фильме о войне во Вьетнаме «В поисках Джека».

**А как насчет мелких подделок?**

Придуманные Сэмом Грегори из правозащитной организации Witness, мелкие подделки - это видео, которые либо представлены без контекста, либо обработаны простыми инструментами редактирования. Они грубые, но, несомненно, действенные.

Например: мелкий фальшивый видеоролик, который замедлил речь Нэнси Пелоси и заставил спикера Палаты представителей США звучать невнятно, распространился по миллионам людей в социальных сетях.

В другом случае корреспонденту CNN Джиму Акосте временно запретили участвовать в брифингах для прессы в Белом доме во время жаркого разговора с президентом. Позже было опубликовано небольшое поддельное видео как он вступает в контакт с практикантом, который пытался отобрать у него микрофон. Позже выяснилось, что видео было, поддельным.

Консервативная партия Великобритании использовала аналогичную тактику мелкой подделки. В преддверии выборов консерваторы подделали телеинтервью с депутатом от лейбористской партии, чтобы создать впечатление, что он не может ответить на вопрос о позиции партии.

С дипфейками шалости только увеличиваются. Мир становится все более синтетическим. А эта технология никуда не денется.

**Как проверяется *дипфейк* ?**

Столкнувшись с лавиной  [фейковых новостей](https://www.muyinteresante.es/cultura/arte-cultura/reportaje/fake-news-el-poder-de-las-mentiras-811555407573) и обработанных видео с такими реалистичными результатами, **журналистская задача проверки данных и источников становится как никогда необходимой.**

В дополнение к появлению специализированных СМИ для разоблачения мистификаций и ложной информации, распространяемой в сети, многие редакции включают расширенные механизмы проверки и предлагают курсы и тренинги для своих сотрудников, которые также включают обучение искусственному интеллекту, поскольку эти знания помогают не только обнаруживать мистификации, но также знать пределы манипуляций.

В любом случае, одним из первых шагов по противодействию видео или изображению является проверка источника: **кто это отправил? Это надежно? Проследить путь предполагаемого***дипфейка***, увидеть, где он был впервые опубликован и кто его опубликовал,** - вот некоторые базовые шаги, которые вы всегда должны делать и которые не требуют уникальных знаний, а только здравого смысла.

Проблема в том, что, хотя ложные новости или видео отрицаются, во многих случаях ущерб уже нанесен. Фальшивые новости распространяются намного быстрее, чем настоящие новости. Они имеют на 70 % больше шансов быть воспроизведенными.

Одно из объяснений состоит в том, что то, чем мы делимся в  [социальных сетях,](https://www.muyinteresante.es/tecnologia/articulo/pueden-las-redes-sociales-destrozar-tu-carrera-851500451832) имеет гораздо больше, чем просто эмоции, и кажется, что мы более склонны распространять новости, которые вызывают у нас гнев, как в случае с большинством *фейковых новостей* .

Итак, прежде чем нажимать кнопку «Поделиться», не углубляясь в информацию, подумайте дважды: вы можете способствовать вирусизации фейковых новостей.