ДИПФЕЙК: РАЗВЛЕЧЕНИЕ, ОБРАЗОВАНИЕ, МАНИПУЛЯЦИЯ

А. И. Соловьев

Белорусский государственный университет,
ул. Кальварийская, 9, 220004, г. Минск, Республика Беларусь,
elan2@tut.by

В статье рассмотрено понятие «дипфейк» и описаны варианты использования дипфейков в цифровом медийном пространстве. Проанализированы характеристики, благодаря которым через дипфейки привлекается и удерживается внимание аудитории цифровой среды.

Ключевыя слова: дипфейк; дипфейк как развлечение; дипфейк в образовании; дипфейк в цифровой среде.

DEEPFAKE: ENTERTAINMENT, EDUCATION, MANIPULATION

A. I. Solovyov

Belarusian State University,
9, Kalvariyskaya, 220004, Minsk, Republic of Belarus
Corresponding author: A. I. Solovyov (elan2@tut.by)

The article discusses the concept of «deepfake» and describes the options for using deepfakes in the digital media space. The characteristics due to which the attention of the audience of the digital environment is attracted and retained through deepfakes are analyzed.

Key words: deepfake; deepfake as entertainment; deepfake in education; deepfake in the digital environment.

Понятие «дипфейк» происходит от английского словосочетания deep fake, которое состоит из слов deep (глубокий) и fake (подделка, фальшивка). Соответственно «глубокая подделка» подразумевает ненастоящие изображения, аудио- и видеоматериалы, сделанные с помощью искусственного интеллекта, когда лицо (или голос) заменяется на лицо (голос) другого человека. Подобные артефакты производились и раньше,

но они требовали кропотливых усилий целой команды фотографов, дизайнеров, специалистов по аудио- / видеомонтажу и 3D-моделированию. Сегодня всю рутинную работу на себя берет искусственный интеллект, производителю контента остается только подготовить исходные файлы и контролировать процесс создания с помощью алгоритмов на основе реальных образцов голоса, звука, видео или снимков, которые сшиваются вместе, когда алгоритмы собирают необходимую информацию из разных источников и затем объединяют ее. В результате возникает нечто новое, искусственно созданное, ненастоящее, но основанное на сочетании реальных данных.

Дипфейком также может быть «снимок» или «видео» с нуля, например, при создании лица или визуального образа персонажа, которого никогда не существовало, но который можно использовать для образования и развлечения. Целью подобных действий, среди прочего, может быть также распространение дезинформации, т. е. умышленная трансляция поддельной информации для намеренного манипулирования (например, о политических и иных деятелях или событиях), а также для различных злоупотреблений (например, финансовых краж через подделку голосовых команд).

Первый способ создания дипфейков предполагает использование двух алгоритмов. Сначала один алгоритм собирает общие черты двух изображений, затем другой алгоритм переносит их на новое созданное изображение. Например, если необходимо создать видео, в котором персонаж двигался бы определенным образом, то алгоритм заменит имеющееся лицо на лицо данного персонажа и заставит его воспроизводить нужные движения. При этом выбранные действия (например, движения, жесты, мимика) копируются и переносятся на новое видеоизображение.

Второй способ создания дипфейков – с помощью генеративно-состязательных сетей (GAN) – также предполагает использование двух алгоритмов. Один из них – генератор – создает образы (например, образ человека с его отличительными признаками: телом, лицом, глазами и т. д.). Другой алгоритм – дискриминатор – оценивает, являются ли изображения, предоставленные ему генератором, правдивыми или нет.

Время появления современных дипфейков – 2017 год. С тех пор технологии получили стремительное развитие, а из-за того, что программы используются в основном в развлекательных целях, их создание не ограничено законом. Многие эксперты считают, что в будущем с развитием технологий дипфейки станут гораздо сложнее и будут нести серьезные

угрозы обществу, например, в связи с вмешательством в выборы, созданием политического напряжения или криминальной деятельностью.

В последние годы технология дипфейка развилась настолько, что сейчас становится все труднее определить, сфабрикованное ли это видео или настоящая запись реальных людей.

Еще одной потенциальной угрозой является контент для взрослых, который всегда генерировал наибольший трафик в виртуальном пространстве. Согласно отчету нидерландской компании Deeptrace, занимающейся кибербезопасностью, 96 % сфабрикованных видео, созданных с помощью технологии дипфейка, являются контентом на такие темы. Чаще всего для изготовления подобных материалов используются образы кинозвезд, спортсменов и даже политиков и представителей власти [1].

Чтобы распознать дипфейк при просмотре видео, стоит обратить внимание на следующее: отсутствие синхронности звука и движений рта; различные явления, которые кажутся неестественными (например, положение головы по отношению к туловищу, неправильное отражение света на предметах, неестественный цвет кожи и т. п.); разница в качестве аудиозаписи и видеосъемки; неровности в изображении, особенно на стыке тела и головы, которая кажется «приклеенной» к другому телу, размытости в области шеи, пробелы кадров (прерывность) или ошибки кадров (различный угол, тип или направление света).

Дипфейк можно найти везде, где существует обширная аудитория, например, на сайтах социальных сетей присутствие «доработанных» видео, фотографий или аудиозаписей почти считается уже нормой. На этой технологии основаны многие развлекательные приложения для смартфонов, позволяющие производить различные спецэффекты в отношении лица и тела. В последнее время дипфейки активно используются в любительском и профессиональном кинематографе. Также в образовательных целях, используя технологию дипфейков, «научились воскрешать» ушедших художников, актеров, певцов, музыкантов, писателей, спортсменов, политиков и других знаменитостей. Так, в Музее Сальвадора Дали во Флориде посетителям представляет свои работы «сам художник», с которым можно пообщаться и сфотографироваться. Технологии дипфейка стали использовать в голосовых генераторах – устройствах для людей, которые потеряли способность говорить.

Сферой негативного использования дипфейков в наше время рассматривается политическая коммуникация, когда ложная информация

ставит задачу оказать влияние на фондовый рынок, целые отрасли экономики, крупные промышленные предприятия, дискредитировать публичных деятелей или результаты выборов, повлиять на ход боевых действий в реальной войне. Дипфейки могут содержать ложные заявления, обвинения или создавать образ лидера определенной политической партии, который не соответствует действительности. Это может привести к негативным последствиям, таким как подрыв доверия к политикам и институтам власти, ухудшение отношений между странами, а также нарушение свободы слова и информации. Хорошо подготовленная поддельная запись облегчает манипулирование общественным мнением и является инструментом для хакеров, шантажистов и интернет-троллей. Неспособность отличить подлинные материалы от подделок может приводить к падению социального доверия и информационному хаосу.

Выросший уровень угроз со стороны киберпреступников, интернет-мошенничество, деятельность так называемых пранкеров – все это открывшиеся возможности для генерирования фальшивых обращений, просьб о поддержке или предложений о помощи, адресованной «жертвам» современных качественных дипфейков. Потоки ложной информации вынуждают социальные сети бороться с опасными для граждан и всего общества дипфейками через обнаружение поддельных видео и их автоматически удаление. Крупные корпорации инвестируют в создание программного обеспечения, которое способно обнаруживать и нейтрализовывать фейковые материалы.

Дипфейки также используются в СМИ, чтобы создать ложные новости или скомпрометировать определенных людей. Это может привести к распространению дезинформации и подорвать доверие к медийным ресурсам. В связи с этим появляется потребность в обучении журналистов и редакторов тому, как распознавать дипфейки и проверять информацию, прежде чем ее публиковать. Также важно развивать непосредственные технологии для обнаружения дипфейков и предотвращения их распространения в СМИ.

Однако в этой ситуации требуются и правовые нормы, определяющие последствия манипулирования информацией или использования чужого имиджа в подобных целях.