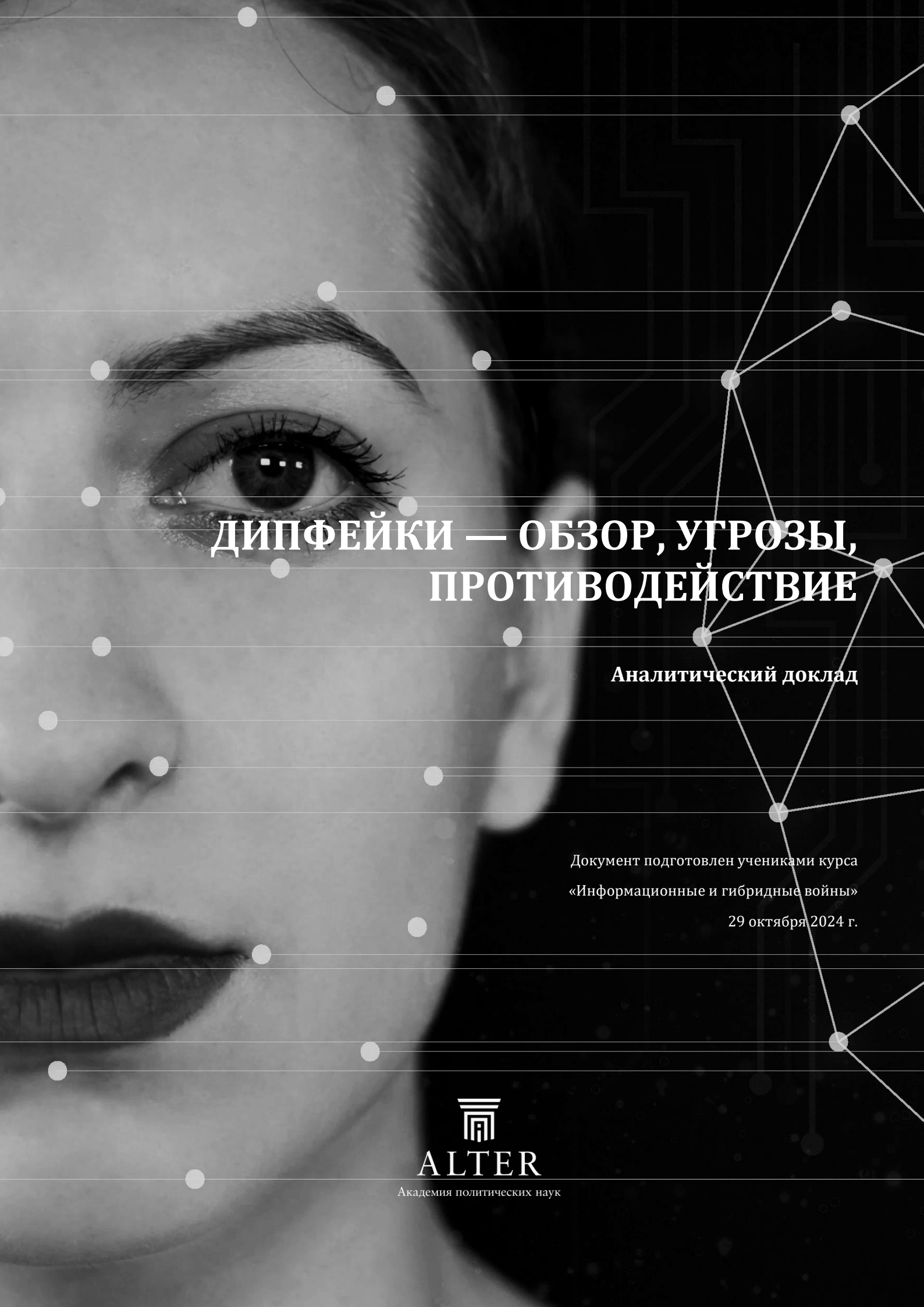

ДИПФЕЙКИ — ОБЗОР, УГРОЗЫ, ПРОТИВОДЕЙСТВИЕ

Аналитический доклад

Документ подготовлен учениками курса
«Информационные и гибридные войны»
29 октября 2024 г.

ALTER
Академия политических наук

Оглавление

Выводы

Дипфейк — довольно широкое понятие, в рамках этого материала рассматривается, как замена части изображения в видео на желаемое (замена лица - faceswap) и подделка голоса — генерация речи с сохраненными характеристиками оригинала. Для изготовления дипфейков используются нейронные сети. В зависимости от используемой программы или модели для изготовления правдоподобного видеоролика может понадобиться от 1 до 10 фотографий «жертвы» (иногда больше). Для синтеза речи может понадобиться от 10 секунд аудио с записью оригинального голоса. Использование технологий синтеза видео и изображений идет рука об руку с другими элементами манипуляции — содержание видеоролика либо звонка будет эмоционально окрашенным, исполнители будут «давить» на срочность (в случае вымогательства) и на другие эмоции (например сострадание, злость, сопричастность и т.д.) в случае атаки на широкую аудиторию. Во время видео или аудио звонка манипулятор будет стараться удержать жертву на линии как можно дольше. И всегда будет содержаться призыв к какому-то действию, так как это есть конечная цель манипулятора. Для технологии «дипфейк» так же могут быть характерными такие признаки: рассинхронизация изображения и звука в видеоролике, движение губ не совпадающее с речью, искажение изображения на границе лица и окружающего фона (появление увеличенных пикселей).

Опасность применения данной технологии кроется в простоте создания дипфейков, что является предпосылкой для массового воздействия на широкую аудиторию, дискредитации и компрометация государственных органов власти, мошенничества с целью получения финансовой выгоды, влияния на обстановку в зоне СВО. На текущий момент, изготовить более-менее простой дипфейк может любой человек, обладающий начальными навыками работы с компьютером. Существует ряд бесплатных программ и приложений, которые используются в развлекательных целях, а так же платных решений, цена за использование которых может начинаться от 20$ в месяц. Если же кто-то обладает чуть более продвинутыми знаниями в информационных технологиях — такому человеку вполне по силам самостоятельно сделать поддельный видео ролик или аудиозапись. При желании изготовление реалистичного дипфейка возможно заказать у фрилансера. Порог вхождения для использования технологии на данный момент крайне низок.

Наиболее реальными исполнителями массированных атак на российское общество могут быть украинские специалисты по информационным

операциям и коллцентры, которые занимаются телефонными мошенничествами. Учитывая продолжительную работу и открытый набор сотрудников, с большой вероятностью можно предполагать, что сеть мошеннических коллцентров контролируется силовыми и криминальными организациями Украины. Это идеальная база из подготовленных специалистов для массового создания и применения дипфейков в отношении российских граждан. Специалисты коллцентров имеют навыки «развода», базирующиеся на знании основ психологии, манипуляции личностью и опытом, совмещенным с отсутствием моральных стопов для занятия преступной деятельностью.

Основными мерами противодействия дипфейкам могут быть: информирование знакомых, родственников, общества о возможностях технологий создания дипфейков, не принимать быстрых решений в состоянии эмоциональной нестабильности.

Виды дипфейков

Подделка/синтез изображения

Обычно используется для подмены лица на фотографии. Для изготовления используются нейронные сети с генеративно-состязательным алгоритмом. Первая сеть генерирует необходимое изображение, а вторая пробует определить, похоже оно на человека или нет. Процесс происходит до тех пор, пока вторая сеть перестанет отличать подделку от реальной фотографии. Можно сказать, что сама по себе замена лица на статическом изображении особой опасности не представляет.

Синтез видео

К технологии, используемой в первом пункте, добавляются детекторы движения. Т.к. видео состоит из последовательности изображений, за процесс все так же отвечают нейронные сети (GAN). Соответственно, возможно заменить оригинальное лицо в видео на лицо другого человека. Технология позволяет:

- генерировать поддельные видео ролики для последующего распространения,

- генерировать поддельные видео сообщения для рассылке через мессенджеры,
- общаться в режиме реального времени по видеосвязи от имени другого человека.

Синтез голоса

На основе оригинального голоса жертвы создается модель нового голоса, который не отличить от образца. Этот голос можно использовать следующим образом:

- создать новую звуковую дорожку для поддельного или оригинального видео (на основе начитанного диктором текста),
- использовать совместно с технологией текст-в-речь (text to speech) и озвучить текст чужим голосом,
- использовать в режиме реального времени для звонков по телефону или в мессенджерах.

Синхронизация движения губ со звуком

Отдельная технология, которая может использоваться совместно с вышеперечисленными — изображение губ на видео меняется в соответствии с произносимым текстом, синхронно, таким образом становится очень сложно отличить подделку от оригинала.

Подготовка

В зависимости от используемой программы или модели для изготовления правдоподобного ролика может понадобится от 1 до 10 фотографий «жертвы» (иногда больше). Для синтеза речи может понадобится от 10 секунд аудио с записью оригинального голоса. Очевидно, что когда большинство людей активно пользуются соцсетями — получить исходных данные для «подделки» личности обычного человека не составит труда. Если же мы говорим о публичных персонах — получить материал еще проще, обычно в сети доступна масса качественных фотографий и выступлений такого человека.

Способы изготовления

На текущий момент, изготовить более-менее простой дипфейк может любой человек, обладающий начальными навыками работы с компьютером. Существует ряд бесплатных программ и приложений, которые используются в развлекательных целях, а так же платных решений, цена за использование которых может начинаться от 20$ в месяц. Если же кто-то обладает чуть более продвинутыми знаниями в информационных технологиях — такому человеку вполне по силам самостоятельно сделать поддельный видео ролик или аудиозапись. При желании изготовление реалистичного дипфейка возможно заказать у фрилансера. Порог вхождения для использования технологии на данный момент крайне низок.

Психологическое воздействие на целевую аудиторию

Важно понимать, что использование технологий синтеза видео и изображений идет рука об руку с другими элементами манипуляции — содержание видео ролика либо звонка будет эмоционально окрашенным, исполнители будут «давить» на срочность (в случае вымогательства), на другие эмоции (например сострадание, злость, сопричастность и т.д.) в случае атаки на широкую аудиторию. В случае видео или аудио звонка манипулятор будет стараться удержать жертву на линии как можно дольше. И всегда будет содержаться призыв к какому-то действию, так как это и есть конечная цель манипулятора.

Дипфейки и Украина

Современные телефонные мошенники из Украины и ЦИПСО могут воздействовать широким инструментарием современных технологий на российское общество. Набирающие популярность дипфейки все больше становятся реальной угрозой кошельку обычных людей и авторитету лиц принимающих решения. Это в зависимости от масштабов атаки может стать вызовом для безопасности на государственном уровне. Учитывая текущую

военно-политическую обстановку, наиболее реальными исполнителями массированных атак на российское общество будут украинские специалисты по информационным операциям и коллцентры, которые занимаются телефонными мошенничествами. Вражеские коллцентры, несмотря на откровенно незаконную деятельность, захлестнувшую не только Россию, но и сопредельные с Украиной страны, редко становятся объектами реальных расследований со стороны правоохранительных органов Украины. Коллцентры крышуются СБУ, депутатами Рады и другими представителями криминальной среды. Инфраструктура коллцентров в виде помещений, оборудования и подготовленных специалистов – идеальная среда для создания мощной сети мошенников, использующих дипфейк технологии. Специалисты коллцентров имеют навыки «развода», базирующиеся на знании основ психологии, манипуляции личностью и опытом, совмещенном с отсутствием моральных стопов для занятия преступной деятельностью. Чтобы выработать меры противодействия такому виду угроз, необходимо понимать основные признаки, задачи и целевые аудитории мошенничеств с применением технологии дипфейка.

Признаки, по которым можно определить дипфейк

Следующие общие признаки могут указывать на попытку манипуляции целевой аудиторией и оказания психологического воздействия:

- упор на эмоции (вызов чувства страха),
- призыв к срочному действию (перевести деньги, выйти на улицу, эвакуироваться, закупаться солью, снимать деньги из банкоматов),
- стремление удерживать контакт со стороны мошенника и остаться единственным источником информации.

Признаки, перечисленные ниже, характерны именно для технологии «дипфейк»:

- нехарактерно короткие аудио и видео сообщения,
- рассинхронизация изображения и звука в видео ролике, движение губ не совпадает с речью,
- искажение изображения на границе лица и окружающего фона (появление увеличенных пикселей),
- неестественное моргание и движение глаз (отсутствие движения),

- аномалии в мимике (эмоции не соответствуют контексту),
- освещение на лице может не совпадать с общим освещением сцены,
- тени могут быть расположены неправильно или отсутствовать,
- движения могут быть дергаными или непропорциональными
- поза тела может оставаться статичной, несмотря на движение головы,
- незначительные изменения цвета или яркости между кадрами, внезапные изменения качества изображения,
- текстуры кожи могут быть сглаженными,
- волосы, зубы или украшения могут выглядеть размытыми,
- голос на аудио может звучать без эмоциональных оттенков,
- интонация может не соответствовать контексту речи,
- отсутствуют естественные паузы для вдоха, переходы между словами неплавные,
- отсутствуют фоновые звуки, характерные для реальной записи.

Целевая аудитория для воздействия, потенциальные мишени

- лица принимающие решения на государственном и региональном уровне,
- непосредственный руководитель организации или учреждения,
- родственники, друзья, знакомые (материалы для создания дипфейка могут браться из соцсетей),
- родственники участников СВО,
- военные в зоне СВО.

Дипфейковые обращения могут идти как из зоны СВО, так и со стороны родственников в адрес военных. Например, военному в зоне СВО могут позвонить или отправить сообщение, от имени жены, рассказав эмоциональную историю о том, что ребенок попал в автомобильную аварию. Сделать подобный вброс могут накануне наступательной операции, ракетного, артиллерийского удара ВСУ, чтобы вывести военного из стабильного эмоционального состояния, в котором он мог бы принимать верные решения.

Цели и задачи создателей дипфейков

1. создание панических настроений в обществе,
2. дискредитация и компрометация государственных органов власти (примером недавнего применения является фейковое обращение губернатора Курской области к гражданам во время первых дней вторжения ВСУ в Курскую область),
3. мошенничество с целью получения финансовой выгоды,
4. шантаж (Например, мошенники начали создавать порноролики с россиянами, чтобы развести их на деньги. Злоумышленники используют нейросети для генерации видео, а после угрожают разослать их родственникам жертвы, если не получат деньги. Сначала мошенники могут написать в WhatsApp, требуя вернуть деньги за несуществующий микрозаём. В случае получения отказа, могут угрожать разослать сгенерированное порно родственникам и друзьям),
5. взлом аккаунтов в соцсетях или на официальных ресурсах (например, Госуслуги),
6. дезинформация о событиях в зоне СВО,
7. влияние на обстановку в зоне СВО,
8. воздействие на лиц принимающих решения.

Меры противодействия

1. информировать знакомых и родственников о возможностях технологий создания дипфейков,
2. не принимать быстрых решений в состоянии эмоциональной нестабильности,
3. если вас просят перечислить деньги или сделать что-то срочно - связаться с этим человеком по альтернативным каналам для уточнения информации,
4. для подтверждения личности человека, задать контрольный вопрос, ответ на который знаете только вы вдвоем или спровоцировать звонящего вопросом, в котором заранее заложена ошибка (например «как поживает твоя жена Наташа?» при условии, что его жену зовут Лена),
5. сделать соцсети закрытыми для широкой аудитории и сократить в них активность,

6. в случае подозрения на мошеннический звонок оборвать разговор, чтобы не дать звонящему получить образец вашего голоса нужной ему продолжительности,
7. в случае поступления информации о тревожных событиях перепроверить информацию в официальных источниках,
8. использовать специальные сервисы, позволяющие определить является ли видео/аудио синтезированным или естественным.