ПОЛИТГЕН
АНАЛИТИЧЕСКИЙ ЦЕНТР

Санкт-Петербург, Лиговский проспект, 74
8 (812) 209 16 50 | info@politgen.ru
www.politgen.ru

Deepfakes: где начинается угроза для личности и национальной безопасности?

Ярослав Игнатовский, политконсультант,
генеральный директор аналитического центра «Политген»

Владимир Иванов, д.полит.н., доцент каф. сравнительной политологии РУДН

В начале февраля широкий резонанс вызвала новость о том, что **индийский политик Маной Тивари эффективно использовал специализированное программное обеспечение «подмены лиц» Deepfake (или Face-Swap)** для того, чтобы создать «дипфейк» собственного рекламного ролика на разных языках и привлечь больше избирателей. Данное событие стало очередной убедительной иллюстрацией как значительного политического и маркетингового потенциала дипфейков, так и возможных угроз, исходящих от их применения, включая манипулирование общественным мнением, вмешательство в личную жизнь граждан и конфликтную мобилизацию этнических или протестных групп под ложными предлогами. Лавинообразное распространение фейковых новостей в политике уже давно вызывает озабоченность и подвергается попыткам государственного регулирования во многих странах. Очевидно, что дальнейшее распространение дипфейков и совершенствование алгоритмов их генерации может привести к дальнейшему снижению доверия к масс-медиа. Доклад Reuters 2019 года показывает, что в мире произошло падение уровня доверия к онлайн-новостям, рост распространения фейковых новостей и снижение доверия к медиаконтенту. Аналогично, Edelman Trust Barometer приходит к выводу, что впервые медиа является наименее доверяемым институтом во всем мире. Такой спад уверенности в «четвертой власти» в первую очередь обусловлен значительным падением доверия к Интернет-платформам, особенно поисковым системам и социальным сетям. Примечательно, что снижение доверия к СМИ сопровождается ростом доверия к информации, рекомендациям и комментариям, размещенным онлайн-пользователями.

Напомним **предысторию скандального прецедента в Индии**: 7 февраля, за день до парламентских выборов в мессенджере WhatsApp набрали популярность два ролика,

в которых глава Индийской народной партии Маной Тивари агитирует голосовать за себя. В одном из видеороликов политик говорил на наиболее распространенном языке хинди, в другом – на диалекте хариани. Однако в действительности, существовал единственный исходник видео, на основе которого и были сделаны дипфейки выступлений политика на других языках. **Для создания ролика** руководство партии привлекло специализирующуюся на политических коммуникациях компанию Ideaz Factory, которая и подготовила дипфейк для «позитивной избирательной кампании», чтобы привлечь избирателей, говорящих на разных языках. Успех данного видео оказался очень значительным: его распространили по 5800 чатам в WhatsApp, а видео посмотрело 15 миллионов человек. После того, как ролик на хариани набрал популярность, компания создала видео на английском языке, чтобы привлечь городских жителей.

Дипфейки индийского политика получили мировой резонанс не только из-за их реальной политической эффективности, но также из-за ряда других факторов: их правдоподобности (так как лицо политика было заменено не полностью, а были сфабрикованы только движения губ, очень немногие пользователи WhatsApp заметили неестественные движения губ кандидата), а также заявления создателей о том, что они впервые использовали технологию дипфейк «на благо», а не для дискредитации оппонентов. Действительно, на первый взгляд, в данной ситуации нет пострадавших, технология использовалась для продвижения позитивного имиджа.

Тем не менее, последний индийский кейс с новой актуальностью поднимает вопрос о том, **насколько легальной и допустимой практикой является использование подобных дипфейков «нового поколения» в публичной, и тем более, в политической деятельности?**

ЧТО ТАКОЕ DEEPFAKES?

Само понятие «Deepfake» образовалось от сочетания терминов **«глубокое обучение» и «подделка».** Deepfakes – это методика компьютерного синтеза изображения, основанная на искусственном интеллекте, которая используется для соединения и наложения существующих изображений и видео на исходные изображения или видеоролики. Искусственный интеллект deepfake использует синтез изображения человека – объединяет несколько картинок, на которых человек запечатлен с разных ракурсов и с разным выражением лица, и делает из них видео. Анализируя фотографии, специальный **алгоритм «самообучается»** тому, как выглядит и может двигаться человек. Сам по себе синтез изображений, видео или аудио может не иметь очевидных социально-опасных целей, однако манипулирование средствами массовой информации с использованием изображений, видео или голосов реальных людей создает целый комплекс моральных и юридических проблем.

Deepfakes может изображать людей, совершающих действия, которых они в действительности никогда не делали, или говорящих такие вещи, которые они никогда не говорили. Формируя модели от сотен до тысяч целевых изображений, алгоритмы deepfake «узнают», как выглядит чье-то лицо под разными углами и в различных выражениях. С помощью самообучения **алгоритм может предсказать**, как будет выглядеть лицо целевого индивида (или жертвы информационной диверсии), имитирующее выражение лица другого человека. Аналогичный процесс используется для тренировки алгоритма deepfake для имитации акцента, интонации и тона чьего-либо голоса.

Технические требования для создания дипфейков невелики. Любой мотивированный человек с ПК среднего уровня может создавать deepfakes. На открытых ресурсах Интернет в свободном доступе находится несколько программ с открытым исходным кодом, например **DeepFaceLab и FaceSwap.**

ПОТЕНЦИАЛЬНЫЕ ОПАСНОСТИ DEEPFAKES

Уже в ближайшем будущем deepfakes может затронуть различные уровни общественной и политической жизни и способствовать распространению широкого спектра угроз: от репутационных рисков для знаменитостей и обычных граждан, до развития организованной преступности и проблем социальной стабильности и национальной безопасности.

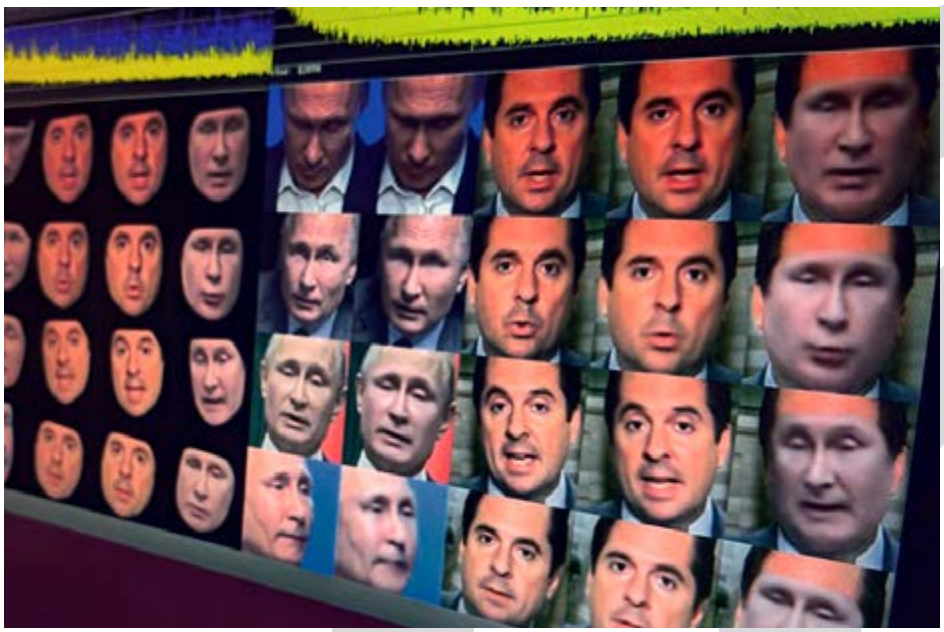
Во-первых, разумеются существуют значительные угрозы злоупотреблений

на индивидуальном уровне. Дипфейки можно использовать для «киберзапугивания», клеветы и шантажа отдельных лиц, в том числе журналистов и политиков. В 2017 году сеть захлестнула волна модной порнографии с deepfake лицами знаменитостей, наложенными на тела порноактрис.

Интернет-травля отдельных публичных лиц (в первую очередь женщин) при помощи дипфейк-контента непристойного и порнографического содержания имеет наиболее разрушительный эффект в странах Азии.

Во-вторых, возможностями технологии deepfake может воспользоваться организованная преступность. Deepfakes может стать золотой жилой для преступных организаций и виртуальных мошенников. Возможно, еще более серьезной проблемой, чем манипуляции с изображениями и видео, является способность технологии имитировать акцент, интонацию и речевые паттерны с невероятной прежде точностью. В то время как многие люди знакомы с возможностями изменять изображения при помощи графических редакторов (например, Photoshop), технология deepfake, особенно голосовая мимикрия, сравнительно неизвестна за пределами областей науки о данных и машинного обучения. Синтезированная речь может использоваться в мошеннических схемах, включая махинации с банковскими счетами, или фиктивные похищениях, когда жертвы сперва отбираются через социальные сети, а потом получают телефонные звонки с требованием выкупа за любимого человека. В цифровую эпоху, когда многие родители открыто делятся видео своих детей в интернете, эта афера может стать значительно опаснее с использованием deepfake audio.

В бизнес-сфере deepfakes можно использовать как наиболее отвратительную форму



черного пиара. Компании или предприниматели смогут заказывать создание deepfake-видео, на которых, например, генеральный директор конкурирующей компании делает клеветнические или оскорбительные заявления, и сливать это видео в социальные сети. Корпоративный саботаж с помощью дипфейков может быть использован и для манипулирования фондовым рынком. Технология может использоваться таким же образом для дискредитации политических оппонентов, партий, общественных движений.

Конечно, **наиболее серьезные опасения** вызывает потенциал использования технологий создания дипфейков для разжигания конфликтов, массовых гражданских беспорядков и подрыва национальной безопасности. Например, во многих странах можно провоцировать межэтнические или межконфессиональные столкновения выкладывая в социальные сети фейковые видео, где представитель той или иной группы высказывает оскорбительные мнения или осуществляет иные действия, которые могут быть восприняты как оскорбление. Если население будет не в курсе о deepfakes и их возможностях, то любое такое поддельное видео с провокационным контентом может приписать любому политику или представителю какого-либо этноса экстремистский посыл. В свою очередь любая попытка властей реагировать и объяснить технологию дипфейков постфактум окажется запоздалой в такой ситуации. По мере развития медиа-технологий их аудитория становится все более вовлеченной. В сочетании с высокой вовлеченностью становится нелегко опровергнуть фальсифицированное видео после того, как оно было просмотрено большим количеством людей.

ВОЗМОЖНОЕ ПРОТИВОДЕЙСТВИЕ РАСПРОСТРАНЕНИЮ ДИПФЕЙКОВ.

Несмотря на реальную опасность deepfakes, следует признать, что данная технология – это в первую очередь просто технический инструмент, который имеет больше положительных применений, чем отрицательных. Однако сегодня правительства стоят перед необходимостью разработать и предпринять определенные действия и меры предосторожности, чтобы свести к минимуму возможность ущерба от использования deepfakes с негативными и преступными намерениями.

Данная проблематика уже стоит в политической повестке ряда стран. Например, летом 2019 года Комитет по разведке Палаты представителей США провел открытые слушания по тематике угроз национальной безопасности, создаваемых искусственным интеллектом, в первую очередь deepfake AI, в ходе которых было единогласно принято решение о том, что deepfakes представляют реальную угрозу для американского общества на различных уровнях. Ведутся дебаты о принятии закона, запрещающего должностным лицам и агентствам Соединенных Штатов создавать и распространять такой контент. В настоящий момент в США идет подготовка проекта соответствующего федерального закона, регулирующего данную сферу.

В России в настоящее время также идет анализ возможностей ограничения неконтролируемого распространения дипфейков в рамках уже принятых законов, направленных на борьбу с недостоверной информацией, публикуемой под видом общественно значимых достоверных сообщений.

Конечно, **распространение опасных дипфейков должно сдерживаться** при помощи внедрения и совершенствования механизмов фактчекинга. Частные лица, социальные медиа-платформы и, особенно СМИ должны иметь инструменты для быстрого и эффективного тестирования информационных сообщений, аудио и видеозаписей, которые они подозревают в подделке. Также желательно, чтобы конечные пользователи – люди могли быть в состоянии определить, является ли подлинной информация, которую они просматривают и которой делятся с другими. Таким образом, приоритетной задачей является развитие сервисов и инструментов фактчекинга (проверки сообщений и выявления фейков). В идеале они должны быть простыми (т.е. не требующими для использования серьезных ИТ навыков и специального образования) и бесплатными.

В качестве приоритетного направления сегодня рассматривается государственный контроль и давление на сервисы социальных сетей с целью более серьезной модерации их контента и внедрения инструментов фактчекинга. Веб-сайты и онлайн-платформы, на которых распространяется опасная фейковая информация, должны нести ответственность и определенную подотчетность. Сегодня анализируются правовые и информационные механизмы, побуждающие социальные сети и мессенджеры более тщательно маркировать «синтетические медиа», повышать осведомленность общественности о таких материалах.

Вопросом времени является введение в действие законов, запрещающих определенный неправомерный контент deepfake. Разработка проектов таких законов уже идет в ряде стран мира.

Параллельно с развитием дипфейк технологий, также совершенствуются технологии их обнаружения и верификации. На данный момент технологии генерации дипфейков еще не смогли полностью преодолеть знаменитый эффект «зловещей долины», согласно которому очень похожий на человека детализированный виртуальный персонаж, вызывает резкую неприязнь и отторжение у аудитории в том случае, если обнаруживаются мелкие несоответствия реальности и даже незначительные движения «как у робота». Видео с применением deepfake пока выглядят убедительно только в первые секунды, чем дольше идет видео, тем сильнее может проявляться эффект «зловещей долины», способный отпугнуть аудиторию и сорвать замыслы манипулятора. Однако для профессиональных провокаторов может оказаться достаточно и нескольких секунд. Специалисты отмечают, что «подрисованные» лица на поддельном видео как правило не моргают, таким образом, в перспективе распознавать дипфейки будет возможно путём анализа движения глаз и частоты моргания.

В то время как человеческий анализ контента необходим, оправдано и создание

автоматизированных инструментов для обнаружения deepfakes. Автоматическое обнаружение может остановить размещение потенциально опасных deepfakes, вместо того чтобы реагировать на такой контент постфактум. Важно, чтобы эти методы были простыми и прикладными, доступными для общества. Помимо обнаружения дипфейков возможна и разработка методов проверки, которые могли бы определить дату, время и физическое происхождение содержимого deepfake.

Таким образом, специалисты по информационной безопасности в разных странах сходятся в том, что для борьбы с распространением общественно опасных дипфейков необходимо повышать осведомленность общественности, развивать технологии обнаружения и внедрять новые законы, регулирующие эту перспективную сферу. Должны разрабатываться и применяться правовые инструменты, что позволит упорядочить данную «серую зону». Однако деятельность, направленная на борьбу с распространением опасного контента, в то же время, не должна подрывать свободу слова.

Такие шаги как обязательная маркировка синтетических медиа, повышенная модерация контента, задержки публикации в социальных сетях и государственное давление на онлайн-платформы для цензуры размещаемого контента неизбежно являются спорными и вызывают неприятие у части общества. Кроме того, под угрозой оказываются бизнес-модели ряда ИТ компаний и информационных ресурсов. Сама природа социальных медиа-платформ предполагает свободу и скорость обмена информацией. В то время как добавление задержек публикации для машинного или ручного анализа контента позволяет отсеивать часть потенциально опасной дезинформации, потеря эффекта мгновенности, даже всего на несколько минут, фактически означала бы изменение сущности социальных медиа в целом. Поэтому сомнительно, что компании, стоящие за популярными социальными сетями, сервисами и месседжерами так просто согласятся на столь радикальное и дорогостоящее изменение своих платформ.

В то же время радикальные меры, подобные инициативам удалить алгоритмы deepfake из публичного доступа являются сомнительными и фактически нереализуемыми. Помимо того, что соответствующее программное обеспечение уже установлено на миллионах компьютеров по всему миру, консервация и игнорирование данной технологии приведет к обратным эффектам – станет намного сложнее противодействовать агрессивной дезинформации с использованием дипфейков, а медиаграмотность, и, соответственно, информационная устойчивость общества искусственно затормозится в своем развитии. Сегодня же происходит постепенный процесс адаптации общества и сетевой культуры к новым медийным возможностям. Дипфейки входят в массовую культуру и эстетизируются, их возможности используются для создания развлекательного контента. **В ближайшие годы мы сможем оценить политический потенциал применения дипфейков и правительствам важно подготовиться к этому.**