RUDN Journal of Public Administration
Вестник РУДН. Серия:
ГОСУДАРСТВЕННОЕ И МУНИЦИПАЛЬНОЕ УПРАВЛЕНИЕ

2020 Том 7 № 4 379–386
http://journals.rudn.ru/
public-administration

DOI: 10.22363/2312-8313-2020-7-4-379-386

Научная статья

Deepfakes: перспективы применения в политике и угрозы для личности и национальной безопасности

В.Г. Иванов

Российский университет дружбы народов
Ул. Миклухо-Маклая, 6, Москва, Россия, 117198

Я.Р. Игнатовский

Аналитический центр «ПолитГен»
Лиговский проспект, 74, Санкт-Петербург, Россия, 191040

Аннотация. В статье анализируется потенциал применения технологий deepfake в политических целях. Отмечается, что 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. Авторами рассматриваются потенциальные угрозы, перспективы и основные направления государственного регулирования данного контента в более широком контексте политических и законодательных инструментов противодействия распространению дезинформации и фейковых новостей.

Ключевые слова: дипфейк, СМИ, политические технологии, информационная безопасность, государственная политика

Введение

В начале февраля 2020 года широкий резонанс вызвала новость о том, что индийский политик Маной Тивари эффективно использовал специализированное программное обеспечение «подмены лиц» Deepfake (или FaceSwap) для того, чтобы создать «дипфейк» собственного рекламного ролика на разных языках для того, чтобы 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. Лавинообразное распространение фейковых новостей в политике уже давно вызывает озабоченность и подвер-

© Иванов В.Г., Игнатовский Я.Р., 2020.
This work is licensed under a Creative Commons Attribution 4.0 International License
https://creativecommons.org/licenses/by/4.0/

гается 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 в разных странах мира [1]. Доклад Reuters 2020 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 масс-медиа во всем мире является институтом, вызывающим наименьшее доверие [2]. Такой глобальный 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 агитировал 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 еще одно 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 кандидата [1].

Тем не менее, рассмотренный 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 син-

теза изображения, основанная на искусственном интеллекте, которая используется для соединения и наложения существующих изображений и видео на исходные изображения или видеоролики. Искусственный интеллект 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, юридических и управленческих проблем.

Deepfakes может достаточно достоверно 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.

Квалификационные и технические требования для создания качественных дипфейков невелики. Любой мотивированный человек с ПК среднего уровня может создавать deepfakes. На открытых ресурсах Интернет в свободном доступе находится ряд 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 [3; 4].

Во-первых, существуют значительные угрозы злоупотреблений на индивидуальном уровне. Дипфейки можно использовать для «киберзапугивания», клеветы и шантажа отдельных лиц, в том числе журналистов и политиков. Так, в 2017 году сеть захлестнула волна 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 азиатских странах.

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, явля-

ется способность технологии имитировать акцент, интонацию и речевые паттерны с недоступной прежде точностью. В то время, как люди преимущественно осведомлены о возможностях 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 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 потенциально использован и для манипулирования фондовым рынком. Таким же образом технология может использоваться для дискредитации политических оппонентов, партий, общественных движений [1].

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 определенной группы высказывается или осуществляет иные действия, которые могут быть восприняты другими как оскорбление. Если широкие слои населения не будут осведомлены о феномене 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 объяснять технологию дипфейков постфактум окажется запоздалой в такой ситуации. По мере развития медиа-технологий их аудитория становится все более вовлеченной. В условиях высокой сетевой вовлеченности становится нелегко опровергнуть фальсифицированное видео после того, как оно было просмотрено большим количеством людей.

Перспективы противодействия распространению дипфейков

Несмотря на реальную опасность deepfakes, следует признать, что данная технология – в первую очередь просто технический инструмент, который имеет больше положительных применений, чем отрицательных. Однако сегодня правительства стоят перед необходимостью разработать и предпринять определенный комплекс действий и мер 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 дипфейки 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 завершается подготовка проекта 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 [1].

Конечно, распространение опасных дипфейков должно сдерживаться при помощи внедрения и совершенствования механизмов фактчекинга (проверки сообщений и выявления фейков). Частные лица, социальные медиаплатформы 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 были 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. В идеале они должны быть простыми (т.е. не требующими для использования серьезных ИТ-навыков и специального образования) и бесплатными, что труднодостижимо и требует соответствующих инвестиций.

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. Таким образом, веб-сайты и онлайн-платформы, на которых распространяется потенциально 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 [5; 6]. 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 неестественные движения «как у робота». Видео с применением deepfake пока выглядят убедительно только в первые

секунды, чем дольше продолжительность 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 путем 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 широкой аудитории. 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 новые 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 [1].

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.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

[1] *Игнатовский Я., Иванов В.* Deepfakes: где начинается угроза для личности и национальной безопасности? // Политген. 05.03.2020. URL: https://www.politgen.ru/analytics/reports/deepfakes-gde-nachinaetsya-ugroza-dlya-lichnosti-i-natsionalnoy-bezopasnosti/. Дата обращения: 02.05.2020.

[2] 2020 Edelman Trust Barometer. 19.01.2020. URL: https://www.edelman.com/trustbarometer. Дата обращения: 02.05.2020.

[3] *Waldrop M.* Synthetic Media: The Real Trouble with Deepfakes // Knowable Magazine. 03.16.2020. URL: https://knowablemagazine.org/article/technology/2020/synthetic-media-real-trouble-deepfakes. Дата обращения: 02.05.2020.

[4] *Fallis D.* The Epistemic Threat of Deepfakes // Philosophy & Technology. 2020. DOI: 10.1007/s13347-020-00419-2

[5] *Tayseer M., Mohammad J., Ababneh M., Al-Zoube A., Elhassan A.* Digital Forensics and Analysis of Deepfake Videos // 11th International Conference on Information and Communication Systems (ICICS). April 2020. DOI: 10.1109/ICICS49469.2020.239493

[6] *Qi H., Guo Q., Juefei-Xu F., Xie X., Ma L., Feng W., Liu Y., Zhao J.* Deep Rhythm: Exposing Deep Fakes with Attentional Visual Heartbeat Rhythms // Proceedings of the 28th ACM International Conference on Multimedia. October 2020. P. 4318–4327. DOI: 10.1145/3394171.3413707

История статьи:
Статья поступила в редакцию: 10.05.2020.
Статья принята к публикации: 15.08.2020.

Research article

Deepfakes: Prospects for Political Use and Threats to the Individual and National Security

V.G. Ivanov

Peoples' Friendship University of Russia (RUDN University)
Miklukho-Maklaya str., 6, Moscow, Russian Federation, 117198

Y.R. Ignatovskiy

Analytical Center PolitGeneration
Ligovskiy prosp., 74, Saint-Petersburg, Russian Federation, 191040

Abstract. The article analyzes the potential of using deepfake technologies for political purposes. It is noted that in the near future, deepfakes may affect various levels of public and political life and contribute to the spread of a wide range of threats: from reputational risks for celebrities and individuals, to the development of organized crime and problems of social stability and national security. The authors evaluate potential threats, prospects and main directions of state regulation of deepfake content in the broader context of political and legislative initiatives to counter the spread of disinformation and fake news.

Keywords: deepfakes, media, political technology, information security, public policy, verification, social media

REFERENCES

[1] Ignatovskiy Y.R., Ivanov V.G. Deepfakes: gde nachinaetsja ugroza dlja lichnosti i nacional'noj bezopasnosti? [Deepfakes: Where Does the Threat to the Individual and National Security Begin?]. *Politgen.* 05.03.2020. URL: https://www.politgen.ru/analytics/reports/deepfakes-gde-nachinaetsya-ugroza-dlya-lichnosti-i-natsionalnoy-bezopasnosti/. Accessed: 02.05.2020 (In Russ.).

[2] *2020 Edelman Trust Barometer.* 19.01.2020. URL: https://www.edelman.com/trustbarometer. Accessed: 02.05.2020.

[3] Waldrop M. Synthetic Media: The Real Trouble with Deepfakes. *Knowable Magazine.* 03.16.2020. URL: https://knowablemagazine.org/article/technology/2020/synthetic-media-real-trouble-deepfakes. Accessed: 02.05.2020.

[4] Fallis D. The Epistemic Threat of Deepfakes. *Philosophy & Technology*. 2020. DOI: 10.1007/s13347-020-00419-2

[5] Tayseer M., Mohammad J., Ababneh M., Al-Zoube A., Elhassan A. Digital Forensics and Analysis of Deepfake Videos. *11th International Conference on Information and Communication Systems (ICICS).* April 2020. DOI: 10.1109/ICICS49469.2020.239493

[6] Qi H., Guo Q., Juefei-Xu F., Xie X., Ma L., Feng W., Liu Y., Zhao J. Deep Rhythm: Exposing Deep Fakes with Attentional Visual Heartbeat Rhythms. *Proceedings of the 28th ACM International Conference on Multimedia.* October 2020: 4318–4327. DOI: 10.1145/3394171.3413707

Article history:
The article was submitted on 10.05.2020.
The article was accepted on 15.08.2020.

Информация об авторах:

Иванов Владимир Геннадьевич – доктор политических наук, доцент кафедры государственного и муниципального управления Российского университета дружбы народов (ORCID ID: 0000-0002-3650-5460) (e-mail: ivanov_vg@pfur.ru).

Игнатовский Ярослав Ринатович – политконсультант, генеральный директор аналитического центра «ПолитГен» (ORCID ID: 0000-0002-2006-4621) (e-mail: hindutime@mail.ru).

Information about the authors:

Vladimir G. Ivanov – Doctor of Political Sciences, Associate Professor of the Department of State and Municipal Management, Peoples' Friendship University of Russia (RUDN University) (Russian Federation) (ORCID ID: 0000-0002-3650-5460) (e-mail: ivanov_vg@pfur.ru).

Yaroslav R. Ignatovskiy – Political Consultant, General Director of the Analytical Center PolitGeneration (Russian Federation) (ORCID ID: 0000-0002-2006-4621) (e-mail: hindutime@mail.ru).

Для цитирования:

Иванов В.Г., Игнатовский Я.Р. Deepfakes: перспективы применения в политике и угрозы для личности и национальной безопасности // Вестник Российского университета дружбы народов. Серия: Государственное и муниципальное управление. 2020. Т. 7. № 4. С. 379–386. DOI: 10.22363/2312-8313-2020-7-4-379-386

For citation:

Ivanov V.G., Ignatovskiy Y.R. Deepfakes: Prospects for Political Use and Threats to the Individual and National Security. *RUDN Journal of Public Administration*. 2020; 7 (4): 379–386. DOI: 10.22363/2312-8313-2020-7-4-379-386