Волкова Александра Евгеньевна

МГУ имени М.В. Ломоносова

**О необходимости регулирования дипфейков**

В последнее время искусственный интеллект (ИИ) распространился на многие сферы нашей жизни и стал их неотъемлемой частью. Несмотря на позитивное влияние этой технологии на некоторые отрасли, существуют опасения, что ИИ может быть использован в недобросовестных целях. Инструментом для этого может стать дипфейк.

Понятие “deepfake” произошло от английских слов “deep” («глубокий»), которое относится к глубокому обучению, и “fake” («подделка») [7]. Дипфейки возникают при сведении пары алгоритмов друг с другом в генеративно-состязательные сети (generative adversarial networks — GANs) и представляют собой аудиовизуальный цифровой контент, который кажется аутентичным, но на самом деле таковым не является [1].

Впервые дипфейки были опубликованы в 2017 году в социальной сети Reddit [3]. Затем появились бесплатные и легкие в применении приложения для генерирования дипфейк-контента, что привело к их распространению [2]. Только к декабрю 2020 года было обнаружено более 85 тысяч вредоносных дипфейковых видеороликов [10]. Сейчас же дипфейки, как и ИИ в целом, получили новый, более качественный виток развития.

В последнее время распространились случаи создания дипфейков в музыкальной индустрии, когда вокал известного исполнителя генерируется с помощью ИИ. Например, песня «Heart on My Sleeve» была создана с помощью дипфейк-голосов известных исполнителей Drake и The Weeknd и даже номинирована на премию «Грэмми» в номинациях «Лучшая рэп-песня» и «Песня года» [9]. Это приводит к распространению дезинформации в интернете, а также к нарушению авторских прав исполнителей, ведь они не давали согласие на использование их голоса. Поэтому среди исследователей ведутся дискуссии о защите неповторимого голоса человека. Как сообщила «Российская газета» в октябре 2023 года, уже началась разработка документа, в котором будет закреплено понятие «синтез голоса» и указана ответственность за создание несогласованной с гражданином аудиодорожки [8].

Глобально дипфейки по цели создания можно разделить на правомерные и противоправные. В группу правомерно созданных дипфейков входит контент, произведенный в развлекательных целях и не нарушающий норм законодательства. Дипфейки, распространяющиеся в противоправных целях, часто выступают как средство для совершения таких правонарушений и преступлений [1], как например:

* Защита чести, достоинства и деловой репутации (ст. 152 ГК РФ);
* Оскорбление (ст. 5.61 КоАП РФ);
* Физическое или психическое принуждение (ст. 40 УК РФ);
* Клевета (ст. 128.1 УК РФ);
* Понуждение к действиям сексуального характера (ст. 133 УК РФ);
* Нарушение авторских и смежных прав (ст. 146 УК РФ);
* Мошенничество (ст. 159 УК РФ);
* Вымогательство (ст. 163 УК РФ);
* Принуждение к совершению сделки или к отказу от ее совершения (ст. 179 УК РФ);
* Преступления против общественной безопасности (некоторые статьи 24 главы УК РФ);
* Незаконные изготовление и оборот порнографических материалов или предметов (ст. 242 УК РФ);
* Преступления против государственной власти (ряд статей X раздела УК РФ);
* Преступления против мира и безопасности человечества (ряд статей 34 главы УК РФ).

В современном медиапространстве дипфейки также могут стать причиной совершения противоправных действий. Это составляет опасность для журналистов: отличить глубокий фейк от подлинного контента становится все сложнее, что может привести к распространению заведомо недостоверной общественно значимой информации под видом достоверных сообщений (ч. 1 ст. 15.3 Федерального закона от 27 июля 2006 г. № 149-ФЗ «Об информации, информационных технологиях и о защите информации»).

Понимая, что теперь почти каждый человек может создавать убедительные дипфейки, необходимость создания методов обнаружения, контроля и противодействия дипфейкам становится очевидной [4]. Можно выделить две меры противодействия дипфейкам: превентивные и выявляющие [7].

К превентивным относится разработка и составление законодательства для регулирования ИИ в целом, и дипфейков в частности. Так, например, в 2019 году в США был создан Законопроект о борьбе с распространением дезинформации путем ограничения технологии видео-дипфейков, а в 2023 году «Правила регулирования дипфейков» появились в Китае. Также в качестве превентивных мер распространения противоправных дипфейков может служить лицензирование и тестирование крупных моделей ИИ, а также маркировка дипфейк-контента.

К выявляющим мерам можно отнести создание алгоритмов обнаружения дипфейков. Исследователи предлагают несколько моделей — например, свёрточные нейронные сети (convolutional neural network — CNN) [6] и рекуррентные нейронные сети (recurrent neural network) [5].

**Список литературы**

1. Смирнов А. А. "Глубокие фейки". Сущность и оценка потенциального влияния на национальную безопасность // Свободная мысль. – 2019. – №. 5 (1677). – С. 63-84.
2. Chamot F., Geradts Z., Haasdijk E. Deepfake forensics: Cross-manipulation robustness of feedforward-and recurrent convolutional forgery detection methods // Forensic Science International: Digital Investigation. – 2022. – Т. 40. – С. 301374.
3. Chen P., Xu M., Wang X. Detecting Compressed Deepfake Images Using Two-Branch Convolutional Networks with Similarity and Classifier // Symmetry. – 2022. – Т. 14. – №. 12. – С. 2691.
4. Dolhansky B. et al. The deepfake detection challenge (dfdc) dataset //arXiv preprint arXiv:2006.07397. – 2020.
5. Güera D., Delp E. J. Deepfake video detection using recurrent neural networks // 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS). – IEEE, 2018. – С. 1-6.
6. Lamichhane B., Thapa K., Yang S. H. Detection of Image Level Forgery with Various Constraints Using DFDC Full and Sample Datasets // Sensors. – 2022. – Т. 22. – №. 23. – С. 9121.
7. Silva S. H. et al. Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models // Forensic Science International: Synergy. – 2022. – Т. 4. – С. 100217.
8. Российская газета: <https://rg.ru/2023/10/15/s-chuzhogo-golosa.html>
9. Forklog: <https://forklog.com/news/ai/ii-pesnyu-s-vokalom-drejka-nominirovali-na-gremmi>
10. Cybernews: <https://cybernews.com/privacy/report-number-of-expert-crafted-video-deepfakes-double-every-six-months/>