«Регуляторный топор» против фейков

Вне зависимости от страны регулирование ИИ и дипфейков — это общая глобальная задача, требующая баланса между развитием технологий, защитой граждан и интересами государства

Текст
ВАДИМ ФЕРЕНЕЦ,
ОБОЗРЕВАТЕЛЬ «Б.О»

В 2025 году компания MWS AI (центр компетенций по ИИ внутри экосистемы МТС) провела эксперимент: 1,6 тыс. взрослых показали 10 фотографий, четыре из которых были сгенерированы нейросетью. Более 60% участников не смогли отличить дипфейки от реальных снимков, а один из поддельных кадров признали настоящим 81% опрошенных.

Чуть ранее MTS AI и Б1 опросили 39 представителей бизнеса из разных отраслей, большинство из них — руководители крупных компаний. Выяснилось, что 92% респондентов считают спуфинг и дипфейки реальной угрозой для бизнеса, а 21% компаний уже пострадали от атак с использованием дипфейков.

Эти и множество других данных привели выступавшие 13 мая 2026 года на «Открытом заседании межведомственной рабочей группы "ДИПФЕЙК"» под руководством **Александра Шойтова,** заместителя министра Минцифры.

От исследований к конкретике

За сухими цифрами, как выясняется, скрываются незаурядного ума злоумышленники, готовые на самые дерзкие преступления, в том числе в области финансового сектора.

Например, 34-летний житель Амстердама с помощью технологий дипфейка смог открыть 46 банковских счетов, используя чужие персональные данные. Для этого он собирал сканы паспортов у людей, которые хотели арендовать у него квартиру. Получив документы, злоумышленник с помощью ИИ изменял свои фотографии: корректировал черты лица — глаза, нос, рот — чтобы они совпадали с данными из документов, либо переносил элементы внешности с чужого паспорта на свое изображение.

Таким образом, он успешно проходил банковскую верификацию по селфи, открывая счета на чужие имена. Злоумышленника задержали в конце 2025 года после случайной проверки на границе, когда сотрудники обнаружили у него подозрительно большое количество банковских карт.

Другой пример: в 2020 году злоумышленники с помощью имитации голоса одного из руководителей банка в ОАЭ убедили сотрудников перевести средства. Ущерб составил 35 млн долларов. Использование аудиоспуфинга (подмена голоса) позволило обойти стандартные процедуры подтверждения операций в банке.

Наконец, довольно распространенная механика: мошенники используют синтезированный голос, чтобы позвонить клиентам банков и представиться их родственниками, попавшими в беду. В результате жертвы переводят деньги на «безопасные счета» мошенников. Дипфейк-голоса и психологическое давление делают схему очень эффективной, а потому крайне опасной.

Позиция МВД

По данным МВД, в 2025 году в России зарегистрировано 1,8 млн преступлений, из которых 690 тыс. (38,2%) связаны с мошенничеством в интернете, в том числе с использованием дипфейков и ИИ.

При этом были использованы следующие виды дипфейков:

- ***визуальные*** — замена лиц, мимики, эмоций, динамические маски в реальном времени. Используются для имитации сотрудников банков, ФСБ, полиции, чтобы вызвать доверие и заставить перевести деньги или взять кредиты;
- ***акустические*** — клонирование и синтез голоса. Для создания достаточно 10–30 секунд записи. Применяются для звонков от имени знакомых, руководителей, родственников;
- ***текстовые*** — генерация сообщений, имитирующих стиль жертвы, для рассылки знакомым с просьбой о финансовой помощи;
- ***мультимодальные*** — сочетание визуальных и акустических технологий для достижения максимального эффекта.

Постепенно кража денежных средств и оформление кредитов на жертв дополняются технологиями формирования общественного мнения и провокациями беспорядков (пример — инцидент в аэропорту Махачкалы). А главное, наблюдается кризис доверия к цифровым доказательствам, осложняющий судопроизводство по подобным делам.

В качестве противодействия дипфейкам в МВД предлагают следующие меры:

- внедрение технологий детектирования дипфейков в мессенджеры и соцсети;
- маркировка контента на уровне платформ, а не авторов;
- легковесные антидипфейк-решения для смартфонов;
- внесение изменений в законодательство по ИБ и особый порядок удаления фейкового контента.

Что думают в Госдуме?

От имени законодательной ветви власти **Антон Горелкин,** первый заместитель председателя комитета Госдумы по информационной политике, информационным технологиям и связи, председатель правления Регионального общественного центра интернет-технологий (РОЦИТ), высказался о маркировке генеративного контента.

Прежде всего нецелесообразно вносить регулирование по фейкам и дипфейкам в Закон № 149-ФЗ «Об информации, информационных технологиях и о защите информации», так как этот закон уже «перегружен».

Распространение фейков сравнивается с эпидемией — пока большинство граждан не научатся распознавать фейки, проблема будет только нарастать. Необходима некая «прививка» цифровой грамотности.

Что касается маркировки контента, то, по словам Антона Горелкина, в Китае и Европе уже внедряется обязательная маркировка ИИ-контента, в США преобладает модель саморегулирования платформ при точечном законодательном вмешательстве. Отечественные платформы пока не дают пользователям инструментов для маркировки или жалоб на фейки, поскольку опасаются неравных условий с иностранными сервисами и дополнительных расходов, но при этом не проявляют инициативы по саморегулированию.

В итоге в Госдуме склоняются к маркировке органического контента, чтобы пользователи могли отличать его от сгенерированного ИИ. Однако такой подход требует проработки механизмов верификации. Важно, чтобы платформы действовали проактивно, а не ждали «регуляторного топора».

Модели регулирования ИИ

Китай — жесткая административная модель, обязательная авторизация пользователей, акцент на маркировке синтетического контента, ответственность по всей технологической цепочке (поставщики, пользователи, платформы). С 2025 года — явная и неявная маркировка, запрет на удаление или подделку маркировки.

Южная Корея — акцент на прозрачности и безопасности, обязательная маркировка для всех систем генерации, управление рисками, запрет на дипфейки в предвыборной кампании.

ЕС — комплексная система: обязательная маркировка, водяные знаки, исключения для сатиры и искусства, оценка рисков, ответственность провайдеров, раскрытие политики по маркировке.

США — минимизация законодательных запретов, регулирование только в чувствительных сферах (интимные дипфейки, выборы, цифровые копии актеров), поддержка инноваций, но делаются попытки создать единую федеральную законодательную рамку.

Что думают юристы?

Представители МГЮА имени Кутафина предлагают следующую модель регулирования для нашей страны:

- закрепить распределение ответственности между провайдерами, пользователями и платформами;
- ввести двухуровневую маркировку (видимую и машиночитаемую);
- запретить удаление/подделку маркировки с административной ответственностью;
- установить специальный режим для дипфейков в избирательном процессе;
- защитить цифровой образ и голос человека на законодательном уровне.

В итоге маркировка контента должна прослеживаться через весь технологический цикл, но не всякий сгенерированный контент — противоправный. Важно разделять маркировку и противоправное использование.

Точку в дискуссии поставил Александр Шойтов: «Проблемы регулирования ИИ и дипфейков схожи во всех странах, подходы динамично меняются. В США также нет единой федеральной рамки, законы разрознены по штатам, что мешает крупному бизнесу. Сейчас обсуждается создание единого закона, но идут споры о детализации и национальной безопасности. Все страны сталкиваются с необходимостью балансировать между инновациями, безопасностью и правами граждан. Россия не уникальна в своих задачах — все ищут оптимальную рамку регулирования». Б.О