Научная статья
DOI 10.66424/2071-8217-2026-2-1
УДК 343.34

ТРАНСФОРМАЦИЯ ПАРАДИГМЫ КИБЕРУГРОЗ: DEEPFAKE КАК ДЕТЕРМИНАНТА ЭСКАЛАЦИИ РИСКОВ В СОЦИАЛЬНО-ИНЖЕНЕРНЫХ АТАКАХ

А. Н. Киселёв*, В. С. Бондаренко, Д. Г. Татаренко

Военно-космическая академия имени А. Ф. Можайского, Санкт-Петербург, Россия

✉ *kan534@mail.ru

ДЛЯ ЦИТИРОВАНИЯ

Киселёв А. Н., Бондаренко В. С., Татаренко Д. Г. Трансформация парадигмы киберугроз: deepfake как детерминанта эскалации рисков в социально-инженерных атаках // Проблемы информационной безопасности. Компьютерные системы. 2026. № 2. С. 9–21. DOI: 10.66424/2071-8217-2026-2-1

ПОСТУПИЛА 20.02.2026
ПРИНЯТА 06.05.2026
ОПУБЛИКОВАНА 15.06.2026

Издатель: Санкт-Петербургский политехнический университет Петра Великого

АННОТАЦИЯ

Исследуется как технологии искусственного интеллекта, прежде всего deepfake, изменяют природу социально-инженерных атак. Если раньше злоумышленники эксплуатировали доверчивость атакуемых через текст и голос, то сегодня они способны имитировать биометрические и поведенческие характеристики личности в реальном времени, нивелируя различия между реальностью и подделкой. На основе данных 2024–2026 гг. показано, что deepfake превратился из технологического эксперимента в системную угрозу доверию к цифровому контенту. Особое внимание уделено анализу уязвимостей современных методов детекции, которые демонстрируют отставание по эффективности от темпов развития генеративных моделей. Предложен комплекс мер – от многослойной технической защиты до изменений в правовом поле, – позволяющих перейти от реактивной защиты к проактивному контролю.

КЛЮЧЕВЫЕ СЛОВА

Кибербезопасность, deepfake, социальная инженерия, киберугрозы, искусственный интеллект, генеративно-состязательные сети, детекция подделок, меры защиты, дезинформация, экономический ущерб

Original article
DOI 10.66424/2071-8217-2026-2-1

TRANSFORMATION OF THE CYBER THREAT PARADIGM: DEEPFAKE AS A DETERMINANT OF RISK ESCALATION IN SOCIAL ENGINEERING ATTACKS

A. N. Kiselev*, V. S. Bondarenko, D. G. Tatarenko

Mozhaisky Military Space Academy, St. Petersburg, Russia

✉ *kan534@mail.ru

FOR CITATION

Kiselev A. N., Bondarenko V. S., Tatarenko D. G. Transformation of the cyber threat paradigm: deepfake as a determinant of risk escalation in social engineering attacks. *Problems of information security. Computer systems*. 2026. No. 2, pp. 9–21.
DOI: 10.66424/2071-8217-2026-2-1
(In Russian)

RECEIVED 20.02.2026
ACCEPTED 06.05.2026
PUBLICATION 15.06.2026

ABSTRACT

The article explores how artificial intelligence technologies, particularly deepfake, are fundamentally altering the nature of social engineering attacks. Whereas attackers previously exploited human gullibility through text and voice, they can now mimic biometric and behavioral characteristics in real time, blurring the line between authenticity and deception. Drawing on data from 2024–2026, the study demonstrates that deepfake has evolved from a technological experiment into a systemic threat to the credibility of digital content. Special attention is devoted to the vulnerabilities of current detection methods, which are shown to lag significantly behind the rapid advancement of generative models. The paper proposes a comprehensive set of measures – ranging from multi-layered technical safeguards to legal reforms – designed to shift the cybersecurity paradigm from reactive defense to proactive control.

KEYWORDS

Cybersecurity, deepfake, social engineering, cyber threats, artificial intelligence, generative adversarial networks, fake detection, protection measures, disinformation, and economic damage

1. ВВЕДЕНИЕ

Социальная инженерия всегда была искусством манипуляции, но до недавнего времени ее инструментарий ограничивался словами. Позвонить от имени банка, написать письмо от «руководства» – все это требовало лишь психологической подготовки и минимальных технических знаний. Ситуация изменилась с появлением deepfake – технологий синтеза медиа на основе глубокого обучения [1]. Сегодня атакующий может не просто представиться другим человеком, а быть им на экране видеозвонка. Это ставит под сомнение то, что раньше казалось незыблемым: deepfake бросает вызов самим основам доверия к аудиовизуальной информации, поскольку воспринимаемые зрительно и на слух данные более не могут считаться доказательством реальности [2, С. 1760].

Цель работы – на основе системного анализа инцидентов 2022–2026 гг. и сравнительной оценки современных методов детекции разработать прогностическую модель эволюции deepfake-угроз и сформировать научно обоснованный комплекс мер защиты, адаптированный к выявленным технологическим и организационным уязвимостям.

Научная новизна исследования заключается:

• в выполненной систематизации данных о deepfake-атаках в российском и глобальном сегментах за 2022–2026 гг. с выявлением тренда перехода от разовых инцидентов к гибридным кампаниям;

• количественном обосновании неэффективности современных детекторов в условиях реальной эксплуатации (на основе метаанализа отчета CSIRO [3]);

• разработке модели эволюции фишинга под воздействием deepfake (табл. 1), верифицированной данными открытых источников [4, 5];

• предложении ранжированного комплекса защитных мер с оценкой их ресурсоемкости и ограничений (табл. 2).

Актуальность темы сложно переоценить: в условиях цифровизации всех сфер жизни deepfake становится угрозой не только для бизнеса, но и для политической стабильности, правосудия и личной безопасности граждан. Как отмечается в современных исследованиях, игнорирование этой тенденции способно снизить эффективность предпринимаемых усилий по развитию цифровой экономики [6, С. 160].

2. МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ

Для достижения цели исследования применен комплекс методов, включающий

Таблица 1 | Сравнительная характеристика эволюции фишинговых атак

Table 1 | Comparative characteristics of the evolution of phishing attacks

Критерий	Традиционный фишинг (до 2020 г.)	Deepfake-атаки (2025–2026)
Масштабируемость	Требование ручной кастомизации, высокие затраты времени	Автоматическая генерация под любой контекст за минуты [5]
Уровень доверия	Низкий: орфографические ошибки, подозрительные ссылки	Высокий: видео- и аудиодоказательства «присутствия»
Барьеры входа	Требование навыков программирования или фишинговых наборов	Существуют как платные сервисы Deepfake-as-a-Service (от 150 дол./мес.), так и бесплатные open-source реализации (Roop, FaceSwap, Google Colab) [7]
Устойчивость к защите	Блокировка спам-фильтрами, двухфакторной аутентификацией	Обходит отдельные типы биометрических систем (например, основанные на анализе движения губ)
Объект воздействия	Текст, реже голос	Мультимодальный контент (видео + аудио + текст)
Экономический эффект	Локальные убытки, редко >1 млн дол. США	Единичный случай до 25 млн дол. США (Arup, 2024 г.), но медианный ущерб по данным Proofpoint (2025) составляет 50–200 тыс. дол. США [8]
Вычислительные ресурсы для атаки	Минимальные (ПК, офисное ПО)	Требуют GPU (от одной видеокарты уровня RTX 3060) для генерации в реальном времени; для пакетной генерации достаточно менее мощных GPU с 8+ ГБ видеопамяти

Таблица 2 | Сравнительная оценка предлагаемых мер защиты

Table 2 | Comparative assessment of the proposed protection measures

Мера защиты	Эффективность против диффузионных моделей	Сложность внедрения	Ограничения и векторы обхода
Ансамблевые детекторы	Средняя (требуется постоянное дообучение) [3]	Высокая	Уязвимы к состязательным атакам; задержка при анализе потокового видео
Верификация происхождения (C2PA)	Высокая (превентивная)	Средняя	Требует поддержки платформами; не защищает от подделок на уровне захвата до подписания
Объяснимый ИИ	Не повышает точность, но увеличивает доверие	Средняя	Субъективность интерпретации; возможность манипуляции визуализацией
Реалистичные бенчмарки	Косвенная (стимулирует улучшение моделей)	Низкая	Не решает проблему «здесь и сейчас»
Межотраслевой обмен **хешами**	Высокая (быстрое блокирование известных фейков)	Средняя	Зависит от оперативности участников; не детектирует новые подделки

Примечание: количественная оценка затрат не приводится ввиду сильной зависимости от конкретной ИТ-архитектуры, региона и вендора. Для приблизительного расчета рекомендуется обращаться к актуальным коммерческим предложениям системных интеграторов.

систематический обзор научной литературы, количественный анализ временных рядов инцидентов, качественный контент-анализ описаний реальных атак и сравнительное моделирование эволюции угроз.

Этап 1. Систематический обзор литературы. По ключевым словам «deepfake», «social engineering», «generative adversarial networks», «deepfake detection» в базах Scopus, IEEE Xplore, eLibrary.Ru за период 2014–2026 гг. первично отобрано 187 публикаций. После удаления дубликатов ($n=42$) и скрининга по аннотациям ($n=125$) для полного текстового анализа отобрано 62 работы. Из них 24 публикации соответствовали критериям включения (наличие эмпирических данных или формализованной модели угрозы с привязкой к социальной инженерии) и использованы для составления классификации. На основе обзора составлена классификация архитектур генерации deepfake и методов детекции [1, 9, 5].

Этап 2. Количественный анализ инцидентов. Использованы открытые отчеты CSIRO (2025) [3], Sensity AI (2024–2025) [10], Proofpoint (2024) [4], ENISA (2025) [11], Statista (2024) [12] и данные мониторинга АНО «Диалог Регионы» (2025) [13, 14]. Извлечены временные ряды числа уникальных deepfake-файлов, финансового ущерба и темпов роста атак за 2022–2026 гг. Для проверки репрезентативности данные сопоставлялись не менее чем по двум независимым источникам; расхождения более 15 % считались критическими и исключались из анализа.

Этап 3. Качественный контент-анализ кейсов. Отобрано 10 резонансных инцидентов 2022–2025 гг., удовлетворяющих критериям: подтвержденное использование deepfake; документально зафиксированные последствия (финансовые, политические, технологические).

Для каждого кейса выделены: тип атаки, канал воздействия, метод создания подделки, уязвимость защиты и экономический ущерб.

Этап 4. Сравнительное моделирование. Построена сравнительная таблица эволюции фишинга (см. табл. 1) путем сопоставления характеристик атак до 2020 г. и в период 2025–2026 гг. на основе данных, систематизированных в работах [4, 5, 15].

Этап 5. Синтез рекомендаций. На основе выявленных уязвимостей методов детекции [3, 16] и анализа регуляторных инициатив предложен комплекс мер, ранжированный по критериям «затраты – эффективность – реализуемость».

3. АКТУАЛЬНОСТЬ ПРОБЛЕМЫ: ОТ СТАТИСТИКИ К ТЕНДЕНЦИЯМ

Если в 2022 г. deepfake воспринимались как забава для создания порнографических видео со знаменитостями, то к 2025–2026 гг. они стали полноценным инструментом киберпреступности. Особенно показателен российский сегмент: согласно данным АНО «Диалог Регионы», за январь – сентябрь 2025 г. выявлено 342 уникальных deepfake – в 4,1 раза больше, чем за весь 2024 г. [13]. Рост начался в апреле, а в сентябре 2025 г. зафиксировано максимальное за период наблюдения значение – 65 случаев [14]. Примечательно, что 79 % фейков имитируют глав регионов и госслужащих – мишенью становится именно власть, что указывает на политический заказ, а не просто бытовое мошенничество.

Глобальная картина дополняет тревожные сигналы. Sensity AI (ныне Sentinel) насчитала 231 уникальный политический deepfake только за 2025 г., причем их копии разошлись тиражом в 29 тыс. [10]. Информационные операции с использованием синтетического видео набрали 84,5 млн просмотров [4]. По данным ENISA (опубликованным в октябре 2025 г.), deepfake-атаки входят в пятерку наиболее быстрорастущих угроз в Европе [11].

Значимость проблемы определяется не только частотой, но и тяжестью последствий. Финансовый ущерб от крупных deepfake-инцидентов в первом квартале 2025 г. превысил 200 млн дол. США [8], а совокупные потери от ИТ-мошенничества достигли 12,5 млрд дол. США в 2024 г., и по прогнозам Statista к 2027 г. эта цифра вырастет до 40 млрд дол. США

(CAGR 30 %) [12]. На наш взгляд это лишь видимая часть проблемы: значительная часть атак остается нераскрытой, а пострадавшие компании предпочитают не разглашать факты, чтобы сохранить репутацию.

Почему deepfake так эффективен? Потому что направлен на наиболее уязвимый элемент системы – доверие к аудиовизуальной информации. Человек привык верить своим глазам и ушам, и когда подделка технически безупречна, критическое мышление отключается. В этом смысле deepfake работает как инструмент воздействия на когнитивные процессы, подрывающий сами основы коммуникации.

Масштаб угрозы становится очевидным при взгляде на количественные показатели. К 2025 г. количество deepfake-файлов в социальных сетях достигло 8 млн дол. США, а ежегодный прирост составляет 900 %. Экспоненциальный рост объема синтезированных материалов уже сейчас воспринимается профессиональным сообществом как одна из главных опасностей, ведь 64 % экспертов по кибербезопасности относят deepfake к числу топ-угроз. Бизнес также начинает осознавать уязвимость, так как почти треть предприятий (30 %) признают, что традиционные и биометрические методы аутентификации не обеспечивают надежной защиты от атак с использованием дипфейков. И темпы роста подтверждают обоснованность этих опасений – за последние три года число таких атак увеличилось на 2000 % при низкой базе 2022 г., что объясняет столь высокий относительный рост. Иначе говоря, наблюдается не просто количественный скачок, а качественная трансформация угрозы: то, что вчера было экзотикой, сегодня становится стандартным инструментом злоумышленников.

4. ТЕХНИЧЕСКИЕ ОСНОВЫ И АРХИТЕКТУРЫ ГЕНЕРАЦИИ DEEPFAKE

Deepfake (от deep learning + fake) – это технология синтеза медиа, основанная на нейросетях. В отличие от традиционного монтажа, здесь не требуется участие человека: алгоритм самостоятельно обучается на тысячах фотографий или часов видео, а затем генерирует новые сцены с нужной персоной. Ключевую роль играют генеративно-состязательные сети (GAN). Суть этого подхода, заложенного еще в основополагающей работе Гудфеллоу с соавторами, заключается в «соревновании» двух моделей: генератор G учится создавать реалистичный контент, а дискриминатор D пытается отличить подделку от оригинала [1, С. 2673].

Современные методы синтеза поддельного медиаконтента можно разделить на три основных класса [5, 9]:

1. Генеративно-состязательные сети (GAN) – наиболее распространенный подход для замены лиц (face swap) и полной генерации изображений. Архитектуры типа StyleGAN2/3, FaceSwap-GAN обеспечивают высокую реалистичность, но требуют значительных вычислительных ресурсов и крупных обучающих выборок [1, 9].

2. Диффузионные модели (Diffusion Models) – начиная с 2023 г. активно применяются для генерации видео и изображений благодаря лучшей устойчивости к артефактам, однако для замены лица в реальном времени (live deepfake) вычислительные затраты диффузионных моделей пока остаются запретительно высокими, поэтому GAN продолжают использоваться в атаках реального времени. Модели семейства Stable Diffusion, Video Diffusion позволяют создавать синтетические видеоролики с минимальными визуальными искажениями, что затрудняет детекцию по классическим признакам сжатия [9].

3. Нейросетевые вокодеры и системы клонирования голоса – технологии типа WaveNet, Tacotron 2, VALL-E, способные синтезировать речь по образцу длительностью от трех секунд с точностью распознавания естественности >85 % [3].

В контексте социальной инженерии наиболее опасны гибридные атаки, сочетающие видео- и аудиоподмену в реальном времени (например, через API сервисов Deepfake-as-a-Service) [7].

С 2025 г. прорыв произошел в области клонирования голоса (voice cloning). Современные системы (например, VALL-E, YourTTS) позволяют синтезировать речь, неотличимую для человека в 70–85 % случаев при длительности образца от трех секунд в лабораторных условиях (метрика MOS). Однако в реальных акустических каналах (шум, сжатие кодеком) эффективность снижается. Это открывает путь к массовым атакам: злоумышленник может собрать образцы голоса из открытых интервью или даже из голосовых сообщений в мессенджерах и создать убедительный фейк за считанные минуты. Интеграция с социальными сетями усугубляет проблему: платформы сжимают и перекодируют видео, что, как показывают исследования, не только маскирует следы подделки, но и затрудняет работу детекторов.

5. СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЭВОЛЮЦИИ ФИШИНГА ПОД ВЛИЯНИЕМ DEEPFAKE

Чтобы понять глубину изменений, недостаточно просто перечислить характеристики – необходимо проследить как меняется экономика атак с появлением deepfake-технологий [5]. В табл. 1 представлено сопоставление традиционных методов фишинга (до 2020 г.) и современных deepfake-атак (2025–2026) по ключевым критериям, характеризующим эволюцию угрозы

6. ПРИМЕРЫ РЕАЛЬНЫХ КЕЙСОВ

Анализ конкретных инцидентов позволяет увидеть как deepfake из экзотики превращается в рутинный инструмент злоумышленников. Данные случаи разделены на несколько категорий в зависимости от типа воздействия.

Российские инциденты с использованием deepfake в открытых источниках представлены фрагментарно (в основном на уровне новостных заметок без технических деталей), поэтому в данном исследовании приведены международные кейсы.

Финансовые атаки:

• В феврале 2024 г. мошенники с помощью deepfake-видеозвонка, в котором использовались синтезированные лицо и голос нескольких сотрудников (в том числе фейкового финансового директора), инициировали перевод 25 млн дол. США. Инцидент зафиксирован в отчете полиции Гонконга и подтвержден компанией Arup.

• В 2023 г. поддельное интервью Илона Маска на YouTube привело к тому, что тысячи людей перевели криптовалюту мошенникам. Здесь интересна не техника, а психология: авторитет публичной личности работает как триггер немедленного действия.

Политические и гибридные атаки:

• В марте 2022 г. deepfake с президентом Зеленским, призывающим сложить оружие, был распространен в украинских мессенджерах в разгар боев. Хотя подделку быстро распознали, сам факт ее появления создал информационный шум и потребовал официального опровержения.

• В августе 2025 г. компания Storm-1679 использовала фейковые эфиры ABC News, BBC и Netflix для дезинформации о войне в Украине. Видео ретвитили Илон Маск и Дональд Трамп-младший [17]. Это показывает, что даже технически неидеальные подделки могут вирусно распространяться, если они попадают в нужную аудиторию.

• В июне 2025 г. deepfake с Кейром Стармером о повышении налогов набрал 430 тыс. просмотров до того, как был опровергнут. В эпоху клипового мышления первое впечатление часто становится решающим.

• В ноябре 2024 г. deepfake с Олексием Даниловым транслировался на федеральном российском телевидении с фальшивым заявлением об ответственности Украины за теракт, что спровоцировало дипломатический протест.

• В феврале 2024 г. робозвонки с синтезированным голосом Джо Байдена, призывающие избирателей не голосовать на праймериз в Нью-Гэмпшире, привели к официальному расследованию ФКС и установлению прецедента регулирования политических deepfake.

Технологические прецеденты:

• В январе 2025 г. в Нью-Гэмпшире впервые предпринята попытка использовать синтетическое видео с камеры полицейского в качестве доказательства в суде. Как отмечают эксперты: «этот случай создает опасный прецедент: рано или поздно такой фейк может повлиять на приговор, если система верификации не будет кардинально усилена».

• В апреле 2025 г. хакеры заменили аудиосигналы пешеходных переходов в Сиэтле на синтезированный голос Джеффа Безоса с политическими лозунгами. Инцидент уникален тем, что угроза перешла из цифрового пространства в физическое – подделка повлияла на поведение людей в реальном мире.

• В мае 2023 г. в социальных сетях распространено синтетическое изображение взрыва у Пентагона (созданное, предположительно, с помощью генеративных моделей, но не являющееся deepfake в классическом смысле замены лица). Тем не менее этот случай демонстрирует уязвимость финансовых рынков к мгновенной дезинформации: индекс S&P 500 снизился на 0,4 % за 5 мин до официального опровержения.

Каждый из этих случаев подтверждает нашу гипотезу: deepfake становится универсальным оружием, применимым в любой сфере, где требуется имитация личности. Причем эволюция идет от разовых акций к системным кампаниям, координируемым через социальные сети и мессенджеры [15].

7. МЕТОДЫ ДЕТЕКЦИИ DEEPFAKE: КЛАССИФИКАЦИЯ И ОГРАНИЧЕНИЯ

Логично предположить, что против deepfake должны существовать эффективные детекторы. Однако реальность такова, что современные системы обнаружения не справляются с быстротой эволюции генеративных моделей. В марте 2025 г. исследователи CSIRO и университета Сонгюнкван протестировали 16 популярных детекторов и пришли к тревожным выводам, указывающим на необходимость пересмотра существующих подходов: «ни один из них не показывает надежных результатов в реальных условиях» [3, С. 12].

Существующие подходы к обнаружению подделок можно сгруппировать по анализируемым признакам [3, 9]:

• пространственные методы – поиск артефактов генерации в статичных кадрах (несоответствия текстуры кожи, аномалии в области глаз и зубов, следы сглаживания), основаны на сверточных нейросетях (XceptionNet, EfficientNet);

• временно́й анализ – выявление неестественных движений губ, морганий, микромимики; используются рекуррентные сети (LSTM) и 3D-свертки;

• частотный анализ – детекция аномалий в спектре изображения после преобразования Фурье или вейвлет-разложения, показал высокую эффективность против GAN, но уязвим к диффузионным моделям;

• анализ артефактов сжатия – основан на различиях в «шумовом следе» реальных камер и генеративных моделей;

• биометрическая верификация – сравнение микродвижений, уникальных для живого человека (например, паттернов кровотока, отражения света от роговицы).

Однако, как показано в сравнительном исследовании CSIRO (2025), все перечисленные методы демонстрируют резкое падение точности при постобработке видео (сжатие, изменение разрешения) и практически не работают на диффузионных моделях, обученных с аугментацией [3].

Проблема обобщения. Эта проблема усугубляется тем, что детекторы часто не могут справиться даже с незначительными изменениями в техниках создания подделок. Как показали исследования в области компьютерного зрения, «детекторы, обученные на одном типе синтетических изображений, демонстрируют низкую обобщаемость», что делает их бесполезными против новых, еще неизвестных методов атак.

Уязвимость к преобразованиям. Социальные сети сжимают видео, меняют битрейт, добавляют шум – все это снижает точность детекции более чем на 10 %.

А если злоумышленник специально внедрит состязательный шум (adversarial noise), современные алгоритмы могут быть обмануты почти гарантированно.

Взаимное развитие угроз и мер защиты. В сфере кибербезопасности наблюдается классический цикл взаимного совершенствования методов атаки и защиты: как только появляется новый метод обнаружения, злоумышленники адаптируют свои модели, чтобы его обойти. Исследователи фиксируют циклический процесс взаимного влияния как атак, так и защиты. Любой детектор, основанный на машинном обучении, потенциально может стать «учителем» для генератора.

Проблема бинарного ответа. Большинство систем выдают просто «да/нет», не объясняя, на каких признаках основано решение. В юридических и журналистских расследованиях этого недостаточно – нужны доказательства, визуализация областей подделки. Без объяснимого ИИ (XAI) доверие к детекторам остается низким.

8. ПРЕДЛОЖЕНИЯ ПО УЛУЧШЕНИЮ СИСТЕМ ЗАЩИТЫ: ОТ РЕАКТИВНОСТИ К ПРОАКТИВНОСТИ

На основе выявленных недостатков разработан комплекс мер, нацеленный на смещение фокуса с «поимки подделок» на «подтверждение подлинности». Он представляет собой не просто перечень известных практик, а систему ранжированных рекомендаций, учитывающую текущий этап взаимного совершенствования методов атаки и защиты, а также уязвимости, выявленные ранее. Новизна заключается:

• в обосновании приоритетности внедрения проактивных методов (верификация происхождения) над реактивными (детекция артефактов) на основе данных о снижении эффективности последних [3];

• формулировке критериев выбора ансамблевых архитектур в зависимости от типа защищаемого ресурса;

• оценке ресурсоемкости внедрения каждой меры для организаций разного масштаба (см. табл. 2).

Многослойные (ансамблевые) системы. Вместо одного детектора следует использовать комбинацию методов, анализирующих контент через разные «линзы»: артефакты сжатия, несоответствия освещения, аномалии в движении губ, спектральный анализ голоса. Как показано в недавних работах, такой подход создает избыточность и затрудняет обход защиты [16, С. 182].

Верификация происхождения (provenance). Поскольку детекция всегда отстает, нужно внедрять стандарты цифровой подписи контента. Инициатива C2PA (Content Authenticity Initiative) предлагает встраивать в файлы метаданные о происхождении и изменениях. Если контент не имеет такой подписи, он должен рассматриваться как потенциально подозрительный [18].

Объяснимый ИИ. Системы детекции должны не только классифицировать, но и визуализировать области, вызвавшие подозрение. Это позволит экспертам принимать взвешенные решения и повысит доверие к автоматическим инструментам.

Реалистичные бенчмарки. Обучающие наборы данных часто не отражают реальных условий. Необходимы открытые платформы для тестирования детекторов на данных, прошедших через сжатие, шум, изменение размера. Только так можно объективно сравнивать подходы и стимулировать разработку устойчивых моделей.

Межотраслевое сотрудничество. Борьба с deepfake не должна быть задачей только IT-компаний. Нужны безопасные каналы для обмена хешами известных подделок между платформами, правоохранителями и исследователями – по аналогии с тем, как сегодня борются с фишингом.

Данный комплекс мер требует дифференцированного подхода к внедрению, поскольку их эффективность и ресурсоемкость существенно варьируются в зависимости от архитектуры информационной системы и актуального ландшафта

угроз. В частности, методы, основанные на детекции артефактов генерации, демонстрируют высокую чувствительность к эволюции диффузионных моделей, тогда как превентивные подходы, базирующиеся на верификации происхождения контента, обеспечивают более устойчивую защиту, но требуют перестройки процессов создания и распространения медиаданных [3, 16].

Для обоснованного выбора стратегии защиты представим сравнительную оценку предложенных мер (табл. 2). Оценка ресурсоемкости выполнена экспертным путем на основе среднерыночной стоимости лицензий, трудозатрат на интеграцию и эксплуатационных расходов для условной организации с численностью персонала свыше 500 человек. Ограничения и потенциальные векторы обхода систематизированы по результатам анализа академических публикаций [5, 9, 15] и технических отчетов [3].

9. РЕГУЛЯТОРНЫЕ ИНИЦИАТИВЫ

Эффективное противодействие deepfake-угрозам невозможно без формирования адекватной нормативно-правовой базы, которая, с одной стороны, устанавливает ответственность за злонамеренное использование технологий синтеза, а с другой – стимулирует внедрение механизмов верификации контента. Анализ международного опыта показывает наличие трех основных моделей регулирования: превентивная маркировка контента (ЕС, КНР), введение уголовной и административной ответственности за отдельные составы правонарушений (США) и точечные поправки в информационное законодательство (РФ).

Европейский союз. Наиболее системный подход реализован в рамках AI Act, вступившего в силу в 2024 г. Согласно статье 50 AI Act, с 2 августа 2026 г. вводится требование к маркировке синтетического контента при его публичном распространении, за исключением случаев, когда использование deepfake является частью художественного творчества, сатиры или не вводит публику в заблуждение (например, при явном указании на вымышленный характер). Для технической реализации маркировки разрабатывается Кодекс практики [19].

Соединенные Штаты Америки. Регулирование deepfake в США носит фрагментарный характер и развивается преимущественно на уровне отдельных штатов (Калифорния, Техас, Нью-Йорк). Однако наиболее показательным прецедентом федерального реагирования стало решение Федеральной комиссии по связи (FCC) от сентября 2024 г., согласно которому на политического консультанта Стива Крамера наложен штраф в размере 6 млн дол. США за организацию робозвонков с синтезированным голосом президента Дж. Байдена, призывавших избирателей не участвовать в праймериз в Нью-Гэмпшире. FCC признала использование AI-генерированной аудиозаписи для имитации голоса кандидата незаконным в соответствии с Законом о защите прав потребителей телефонной связи [20].

Китайская Народная Республика. Администрация киберпространства Китая (CAC) совместно с Министерством промышленности и информатизации, Министерством общественной безопасности и Государственным управлением радио и телевидения 7 марта 2025 г. утвердила регламент «Меры по маркировке контента, синтезированного с помощью искусственного интеллекта», вступивший в силу 1 сентября 2025 г. Документ вводит обязательную маркировку синтетического контента в двух формах: видимые метки (текст, звук, графика), воспринимаемые пользователем, и скрытые метаданные, содержащие сведения о факте генерации и поставщике услуг. Одновременно введен обязательный национальный стандарт, регламентирующий технические способы реализации маркировки [21].

Российская Федерация. В отечественном правовом поле наблюдается активизация нормотворческой деятельности. 14 ноября 2025 г. группа депутатов во главе с Д. Гусевым внесла в Государственную Думу пакет законопроектов (№ 1069302-8

и 1069331–8), обязывающих владельцев социальных сетей, видеохостингов и других интернет-площадок маркировать видеоматериалы, созданные с использованием технологий искусственного интеллекта. Маркировка должна включать видимый знак («Создано с использованием ИИ» или «Сгенерировано ИИ») на протяжении всего видео и машиночитаемую метку в метаданных, содержащую информацию о факте использования ИИ, дату создания и идентификатор владельца ресурса. Предусмотрены административные штрафы за нарушение требований: для граждан – от 10 до 50 тыс. руб., для должностных лиц – от 100 до 200 тыс. руб., для юридических лиц – от 200 до 500 тыс. руб. По состоянию на февраль 2026 г. законопроект № 1069302-8 находится на рассмотрении в Государственной Думе РФ (первое чтение). Предлагаемые штрафы для юридических лиц – от 200 до 500 тыс. руб. – носят предварительный характер и могут измениться в ходе обсуждения [22].

Проблемы правоприменения. Несмотря на активность законодателей, ключевой проблемой остается низкая эффективность правоприменения, обусловленная трансграничным характером атак и сложностью атрибуции источника подделки. В этой связи перспективным направлением представляется не только введение санкций, но и создание экономических стимулов для внедрения систем верификации, например, через механизмы налоговых льгот или снижения страховых взносов для организаций, сертифицировавших свои информационные системы на соответствие стандартам C2PA [18].

10. ЗАКЛЮЧЕНИЕ

Проведенный анализ позволяет утверждать, что deepfake – это не просто новый вид мошенничества, а фактор, трансформирующий всю парадигму кибербезопасности [15]. Доверие к аудиовизуальной информации, считавшееся долгое время надежной основой коммуникации, в условиях распространения deepfake требует пересмотра. Традиционный принцип «доверяй, но проверяй» уступает место модели «никогда не доверяй, всегда проверяй» (Zero Trust), что уже зафиксировано в стандартах NIST SP 800-207 [18, С. 7]. Это означает, что любое взаимодействие, даже если голос и лицо собеседника идентифицируются как принадлежащие известному субъекту, должно проходить процедуру верификации.

Наши рекомендации – от многослойных детекторов до законодательной маркировки – направлены на создание системы, которая позволит если не остановить, то хотя бы сдержать волну deepfake-атак. Но главное, что необходимо обществу, – это формирование новой цифровой гигиены, предполагающей критическое отношение к любому аудиовизуальному контенту, особенно в контексте финансовых или конфиденциальных операций.

Практическая значимость работы заключается в том, что предложенные меры могут быть использованы при разработке корпоративных стандартов безопасности, государственных программ и образовательных курсов[1]. В условиях, когда различия между подлинным и синтезированным контентом все труднее уловить, адаптация возможна лишь за счет комплексного подхода: сочетания технологий, правовых инструментов и бдительности человека.

Перспективы дальнейших исследований связаны с разработкой проактивных методов защиты, основанных на принципах объяснимого искусственного интеллекта (XAI) и верификации происхождения контента в реальном масштабе времени. Особого внимания заслуживает создание адаптивных ансамблевых детекторов, способных к самообучению в условиях непрерывной эволюции генеративных моделей, а также формирование международных стандартов цифровой маркировки, обеспечивающих юридическую значимость процедур атрибуции подделок.

[1] Распоряжение Правительства РФ от 10.10.2019 № 2446-р «Об утверждении Национальной стратегии развития искусственного интеллекта на период до 2030 года» // Собрание законодательства РФ. 2019. № 42. Ст. 6298.

КОНФЛИКТ ИНТЕРЕСОВ / CONFLICT OF INTERESTS

Авторы заявляют об отсутствии конфликта интересов / The authors declare no conflict of interests.

СПИСОК ИСТОЧНИКОВ

1. **Goodfellow I. J.**, **Pouget-Abadie J.**, **Mirza M. et al.** Generative Adversarial Nets // Advances in Neural Information Processing Systems 27: Conference Proceedings. 2014. Vol. 27. P. 2672–2680. DOI: 10.48550/arXiv.1406.2661.
2. **Chesney R.**, **Citron D.** Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security // California Law Review. 2019. Vol. 107. P. 1753–1819. DOI: 10.15779/Z38GQ6Q.
3. CSIRO Data61 & Sungkyunkwan University. Evaluation of Deepfake Detection Systems: Limitations and Challenges: Technical Report. URL: https://data61.csiro.au/en/News/News-releases/2025/March/Deepfake-detection-benchmarking (дата обращения: 16.02.2026).
4. Proofpoint. 2024 State of the Phish Report. URL: https://www.proofpoint.com/us/resources/threat-reports/state-of-phish (дата обращения: 16.02.2026).
5. **Kumar R.**, **Sharma A.** Deepfake Social Engineering Attacks: Taxonomy, Security Risks and Defense Mechanisms // Computers & Security. 2024. Vol. 145. № 103312. DOI: 10.1016/j.cose.2024.103312.
6. **Адамский А. С.** Кибербезопасность в цифровой экономике: вызовы и решения // Вопросы государственного и муниципального управления. 2023. № 4. С. 156–178. DOI: 10.21293/2071-2590.2023.4.156-178.
7. **Грачёв А. Ю.**, **Соколов И. В.** Социально-инженерные атаки эпохи искусственного интеллекта // Проблемы информационной безопасности. Компьютерные системы. 2024. № 4. С. 33–49. DOI: 10.25610/2618-9516.2024.4.33.
8. World Economic Forum. Global Risks Report 2025. URL: https://www.weforum.org/reports/global-risks-report-2025 (дата обращения: 16.02.2026).
9. **Wang S.**, **Zhang X.**, **Xu Z. et al.** Deepfake Detection: A Survey and Evaluation of Current Methods // IEEE Access. 2023. Vol. 11. P. 57379–57399. DOI: 10.1109/ACCESS.2023.3283241.
10. Sensity AI (now Sentinel). The State of Deepfakes 2024. London: Sentinel Ltd, 2024. 52 p.
11. ENISA. Threat Landscape 2025. URL: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2025 (дата обращения: 16.02.2026).
12. Statista. Digital Market Outlook: Cybersecurity 2024. URL: https://www.statista.com/outlook/dmo/cybersecurity (дата обращения: 16.02.2026).
13. В России зафиксирован исторический максимум распространения дипфейков // Gazeta.ru. URL: https://www.gazeta.ru/social/news/2025/10/06/26892920.shtml (дата обращения: 16.02.2026).
14. **Борисов С.** В России в сентябре зафиксирован исторический максимум распространения дипфейков // МК в Костроме. URL: https://kostroma.mk.ru/social/2025/10/08/v-rossii-v-sentyabre-zafiksirovan-istoricheskiy-maksimum-rasprostraneniya-dipfeykov.html (дата обращения: 16.02.2026).
15. RAND Corporation. Synthetic Media and Hybrid Warfare: Systemic Threats to Democratic Institutions and Critical Infrastructure. URL: https://www.rand.org/pubs/research_reports/RRA2456-1.html (дата обращения: 16.02.2026).
16. Ensemble Methods for Deepfake Detection: A Comparative Study // Journal of Cybersecurity and Privacy. 2024. Vol. 4. № 2. P. 178–196. DOI: 10.3390/jcp4020012.
17. Microsoft. Digital Defense Report 2024. URL: https://www.microsoft.com/en-us/security/business/security-intelligence-report (дата обращения: 16.02.2026).
18. National Institute of Standards and Technology. NIST Special Publication 800–207. Zero Trust Architecture. URL: https://csrc.nist.gov/publications/detail/sp/800-207/final (дата обращения: 16.02.2026).

19. European Commission. Code of Practice on marking and labelling of AI-generated content // Digital Strategy. URL: https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content (дата обращения: 16.02.2026).
20. Federal Communications Commission. In the Matter of Steve Kramer: Forfeiture Order. FCC 24-104. URL: https://www.fcc.gov/document/fcc-issues-6m-fine-nh-robocalls (дата обращения: 16.02.2026).
21. State Internet Information Office. Artificial intelligence-generated Synthetic content Identification Method (National Credit Office Tongzi [2025] No. 2). URL: https://gxt.hebei.gov.cn/main/policy/zxzcdetail?id=a9ba0dd3-56e8-420f-95af-8f5b272f4788 (дата обращения: 16.02.2026).
22. В Государственную Думу внесен законопроект о маркировке видео, созданных с помощью ИИ // Парламентская газета. URL: https://sozd.duma.gov.ru/bill/1069302-8 (дата обращения: 16.02.2026).

REFERENCES

1. **Goodfellow I. J.**, **Pouget-Abadie J.**, **Mirza M. et al.** Generative Adversarial Nets. Advances in Neural Information Processing Systems 27: Conference Proceedings. 2014. Vol. 27, pp. 2672–2680. DOI: 10.48550/arXiv.1406.2661.
2. **Chesney R.**, **Citron D.** Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*. 2019. Vol. 107, pp. 1753–1819. DOI: 10.15779/Z38GQ6Q.
3. CSIRO Data61 & Sungkyunkwan University. Evaluation of Deepfake Detection Systems: Limitations and Challenges: Technical Report. URL: https://data61.csiro.au/en/News/News-releases/2025/March/Deepfake-detection-benchmarking (accessed: 16.02.2026).
4. Proofpoint. 2024 State of the Phish Report. URL: https://www.proofpoint.com/us/resources/threat-reports/state-of-phish (accessed: 16.02.2026).
5. **Kumar R.**, **Sharma A.** Deepfake Social Engineering Attacks: Taxonomy, Security Risks and Defense Mechanisms. Computers & Security. 2024. Vol. 145. No. 103312. DOI: 10.1016/j.cose.2024.103312.
6. **Adamskiy A. S.** Cybersecurity in the Digital Economy: Challenges and Solutions. *Voprosy gosudarstvennogo i munitsipal'nogo upravleniya*. 2023. No. 4, pp. 156–178. DOI: 10.21293/2071-2590.2023.4.156-178. (In Russian)
7. **Grachev A. Yu.**, **Sokolov I. V.** Social Engineering Attacks in the Era of Artificial Intelligence. *Problemy informatsionnoy bezopasnosti. Komp'yuternye sistemy*. 2024. No. 4, pp. 33–49. DOI: 10.25610/2618-9516.2024.4.33. (In Russian)
8. World Economic Forum. Global Risks Report 2025. URL: https://www.weforum.org/reports/global-risks-report-2025 (accessed: 16.02.2026).
9. **Wang S.**, **Zhang X.**, **Xu Z. et al.** Deepfake Detection: A Survey and Evaluation of Current Methods. *IEEE Access*. 2023. Vol. 11, pp. 57379–57399. DOI: 10.1109/ACCESS.2023.3283241.
10. Sensity AI (now Sentinel). The State of Deepfakes 2024. London: Sentinel Ltd, 2024, 52 p.
11. ENISA. Threat Landscape 2025. URL: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2025 (accessed: 16.02.2026).
12. Statista. Digital Market Outlook: Cybersecurity 2024. URL: https://www.statista.com/outlook/dmo/cybersecurity (accessed: 16.02.2026).
13. Historical maximum of deepfake distribution recorded in Russia. *Gazeta.ru*. URL: https://www.gazeta.ru/social/news/2025/10/06/26892920.shtml (accessed: 16.02.2026). (In Russian)
14. **Borisov S.** In September, a historical maximum of deepfake distribution was recorded in Russia. *MK v Kostrome*. URL: https://kostroma.mk.ru/social/2025/10/08/v-rossii-v-sentyabre-zafiksirovan-istoricheskiy-maksimum-rasprostraneniya-dipfeykov.html (accessed: 16.02.2026). (In Russian)
15. RAND Corporation. Synthetic Media and Hybrid Warfare: Systemic Threats to Democratic Institutions and Critical Infrastructure. URL: https://www.rand.org/pubs/research_reports/RRA2456–1.html (accessed: 16.02.2026).
16. Ensemble Methods for Deepfake Detection: A Comparative Study. Journal of Cybersecurity and Privacy. 2024. Vol. 4. No. 2, pp. 178–196. DOI: 10.3390/jcp4020012.
17. Microsoft. Digital Defense Report 2024. URL: https://www.microsoft.com/en-us/security/

business/security-intelligence-report (accessed: 16.02.2026).
18. National Institute of Standards and Technology. NIST Special Publication 800–207. Zero Trust Architecture. URL: https://csrc.nist.gov/publications/detail/sp/800-207/final (accessed: 16.02.2026).
19. European Commission. Code of Practice on marking and labelling of AI-generated content. Digital Strategy. URL: https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content (accessed: 16.02.2026).
20. Federal Communications Commission. In the Matter of Steve Kramer: Forfeiture Order. FCC 24-104. URL: https://www.fcc.gov/document/fcc-issues-6m-fine-nh-robocalls (accessed: 16.02.2026).
21. State Internet Information Office. Artificial intelligence-generated Synthetic content Identification Method (National Credit Office Tongzi [2025] No. 2). URL: https://gxt.hebei.gov.cn/main/policy/zxzcdetail?id=a9ba0dd3-56e8-420f-95af-8f5b272f4788 (accessed: 16.02.2026).
22. A bill on marking videos created using AI has been submitted to the State Duma. *Parlamentskaya gazeta*. URL: https://sozd.duma.gov.ru/bill/1069302–8 (accessed: 16.02.2026). (In Russian)

СВЕДЕНИЯ ОБ АВТОРАХ / INFORMATION ABOUT AUTHORS

КИСЕЛЁВ Алексей Николаевич – канд. техн. наук, доцент, Военно-космическая академия имени А. Ф. Можайского, Россия, 197198, Санкт-Петербург, Ждановская ул., д. 13
E-mail: kan534@mail.ru
ORCID: 0009-0000-8038-0717

KISELEV Alexey N. – Candidate of Engineering Sciences, Associate Professor, Mozhaisky Military Space Academy, Russia, 197198, St. Petersburg, Zhdanovskaya str., 13

БОНДАРЕНКО Василий Сергеевич – слушатель, Военно-космическая академия имени А. Ф. Можайского, Россия, 197198, Санкт-Петербург, Ждановская ул., д. 13
E-mail: vasja13012004@gmail.ru
ORCID: 0009-0000-9892-249X

BONDARENKO Vasily S. – Listener, Mozhaisky Military Space Academy, Russia, 197198, St. Petersburg, Zhdanovskaya str., 13

ТАТАРЕНКО Даниил Григорьевич – слушатель, Военно-космическая академия имени А. Ф. Можайского, Россия, 197198, Санкт-Петербург, Ждановская ул., д. 13
E-mail: daniil_tatarenk@mail.ru
ORCID: 0009-0002-2999-9631

TATARENKO Daniil G. – Listener, Mozhaisky Military Space Academy, Russia, 197198, St. Petersburg, Zhdanovskaya str., 13