АНАЛИТИЧЕСКИЙ ОБЗОР

КОНСОРЦИУМ ИССЛЕДОВАНИЙ БЕЗОПАСНОСТИ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Международный опыт распознавания и маркировки ИИ-контента

Анализ регуляторных подходов США и ЕС, кейсы платформ и методы обхода

13 мая 2026

Докладчик: Служеникин Д.И.

Для межведомственной рабочей группы для проработки проблем противодействия противоправному использованию технологий типа «Deepfake»

КОНСОРЦИУМ ИССЛЕДОВАНИЙ БЕЗОПАСНОСТИ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

01

ПРОБЛЕМА

Эпоха синтетического контента

Стирание граней

68% дипфейков практически неотличимы от реального контента. Технологии генеративного ИИ достигли уровня, где визуальная идентификация стала ненадёжной.

Риски мошенничества

Финансовые потери от дипфейк-мошенничества превысили $200 млн в Q1 2025 только в Северной Америке. Средний ущерб на инцидент — $500K.

Манипуляция сознанием

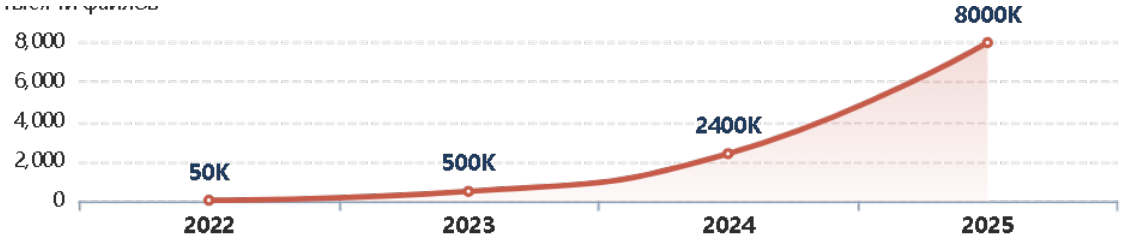
77% избирателей в США столкнулись с дипфейк-контентом, связанным с политическими кандидатами, перед выборами 2024 года.

Экспоненциальный рост дипфейков

2023 → 2025

500K → 8M

Рост на **1500%**



3000%
рост инцидентов в 2023

1740%
рост в Северной Америке

$40B
потери к 2027 (прогноз)

Источник: Sumsub, Pindrop, Deloitte Center for Financial Services, 2024-2025

Калифорнийский подход vs Федеральные инициативы

КОНСОРЦИУМ ИССЛЕДОВАНИЙ БЕЗОПАСНОСТИ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Отсутствие единого федерального закона

США не имеют всеобъемлющего федерального законодательства о дипфейках. Вместо этого действует **фрагментированная система** из 169 законов, принятых 46 штатами с 2022 года.

46	169
штатов с законами	законов с 2022

Фокус на ответственности платформ

Калифорнийские законы возлагают **прямую ответственность** на крупные онлайн-платформы (>1 млн пользователей в штате):

- Идентификация и удаление обманчивого контента за 120 дней до выборов
- Маркировка сообщённого контента в течение 72 часов
- Предоставление механизмов для жалоб пользователей

Калифорния — лидер регулирования

AB 730 (2019): Запрет на распространение дипфейков политических кандидатов за 60 дней до выборов

AB 2355 (2025): Обязательное раскрытие ИИ-контента в политической рекламе

AB 2655: Обязанности платформ по удалению/маркировке дипфейков

AB 2839: Запрет на обманчивый ИИ-контент в избирательных коммуникациях

Проблемы и вызовы

1. **Судебные оспаривания**
   X (Twitter) оспаривает AB 2655 как нарушение Первой поправки
2. **Фрагментированность**
   Разные правила в 46 штатах создают сложности для платформ
3. **Ограниченная эффективность**
   Исключения для сатиры и новостей ослабляют защиту

Ключевой вывод: Калифорния устанавливает стандарты для других штатов, но отсутствие федеральной координации и судебные вызовы ограничивают эффективность регулирования.

03

РЕГУЛЯТОРИКА ЕС

AI Act как глобальный стандарт

AI Act: первый комплексный закон

Европейский акт об ИИ — **первое в мире** комплексное законодательство о регулировании ИИ. Вступает в силу август 2026

Риск-ориентированный подход: ИИ-системы классифицируются по уровням риска — от минимального до неприемлемого

Статья 50: Обязательная маркировка

Провайдеры обязаны:
Обеспечивать машиночитаемую маркировку и детектируемость ИИ-контента

Развёртчики (deployers) должны:
Раскрывать использование ИИ для создания реалистичного синтетического контента

Два подхода к маркировке

Process-based

Маркировка на этапе генерации контента — встраивание метаданных непосредственно в ИИ-системы

Пример: C2PA, водяные знаки

Impact-based

Маркировка в зависимости от потенциального вреда — особые требования для высокорискованного контента

Пример: Политический контент, deepfakes

Code of Practice (2026)

Добровольный инструмент для подготовки к обязательным требованиям AI Act. Включает разработку единого EU-иконки для маркировки

Цифровые сервисы (DSA)

Платформы несут ответственность за распространение незаконного контента и дезинформации, включая дипфейки

Глобальное влияние: AI Act устанавливает стандарты, которые влияют на практику компаний по всему миру ("Брюссельский эффект")

04

ТЕХНОЛОГИЧЕСКИЙ СТАНДАРТ

C2PA: Цифровая подпись контента

Что такое C2PA?

Coalition for Content Provenance and Authenticity — открытый технический стандарт для обеспечения прозрачности и аутентичности цифрового контента.

> "Что если сам контент мог бы рассказать свою историю?"
> — Adobe, основатель инициативы

Как это работает

1. **Provenance (Происхождение)**
   Криптографически защищённая цепочка истории контента от создания до публикации
2. **Метаданные**
   Встраивание информации непосредственно в файлы изображений, видео, аудио
3. **Цифровые подписи**
   Верификация подлинности через криптографические методы

Ключевые участники

Adobe	Microsoft	Google	Meta
Amazon	Intel	Sony	BBC

Примеры внедрения

Adobe CAI: Интеграция в Photoshop, Premiere Pro — 5000+ участников

Microsoft: Поддержка C2PA в продуктах для корпоративных клиентов

Google Pixel 10: Первый смартфон с сертификацией C2PA highest tier

⚠ **Ограничение:** C2PA не защищает от намеренного удаления метаданных — нужны дополнительные меры

КЕЙС 1

05

Успехи и провалы платформ (YouTube, TikTok)

КОНСОРЦИУМ ИССЛЕДОВАНИЙ БЕЗОПАСНОСТИ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Аудит Indicator (2025)

Исследователи загрузили 516 постов с ИИ-контентом на 5 платформ за 3 недели. Результаты показали системные проблемы с маркировкой.

Только

33%

постов получили корректную маркировку

* Для контента от OpenAI Sora

Результаты по платформам

Pinterest	55%
LinkedIn	25%
Instagram	17%
YouTube	~30%
TikTok	0%*

Ключевые проблемы

Несоответствие

Google и Meta не маркируют контент, созданный их собственными ИИ-инструментами

Фрагментарность

TikTok маркирует только контент, созданный внутри приложения

Ложные срабатывания

Системы детекции несовершенны — пропускают или ошибочно маркируют контент

Масштаб задачи

Платформы не справляются с объёмом ИИ-контента — 70% видео на YouTube используют ИИ

YouTube vs TikTok

YouTube лучше справляется с маркировкой видео от Sora, чем TikTok, который не маркирует такой контент вообще

C2PA и платформы

Все audited компании входят в steering committee C2PA, но реализация оставляет желать лучшего

Вывод: Технологии детекции несовершенны, платформы не справляются с масштабом задачи. Необходимы более строгие регуляторные требования.

06

КЕЙС 2

Reddit и борьба с джейлбрейками

Ситуация

r/ChatGPTJailbreak — крупнейшее сообщество Reddit по обмену методами обхода ограничений ИИ (джейлбрейкам). Сообщество накопило сотни техник для "взлома" ChatGPT и других LLM.

100K+	500+
подписчиков	методов джейлбрейка

Действие Reddit

Январь 2026
Полная блокировка сабреддита r/ChatGPTJailbreak

Причина
Нарушение политики платформы относительно вредоносного контента

Цель
Предотвращение распространения методов обхода защит ИИ

Результат

Блокировка привела к устранению среды обмена опытом , но не самой технологии обхода:

- ✓ Методы джейлбрейка остаются доступными через другие каналы
- ✓ Telegram, Discord, отдельные сайты продолжают распространение
- ✓ Новые сообщества быстро создаются для замены заблокированных

Парадокс борьбы с джейлбрейками

Гидра эффект
Блокировка одного сообщества приводит к появлению двух новых

Децентрализация
Информация распространяется через множество каналов, не контролируемых одной платформой

Технологическая гонка
Разработчики ИИ и создатели джейлбрейков находятся в постоянной гонке вооружений

Урок: Блокировка сообществ — временная мера, но не решение проблемы. Необходим комплексный подход: технические меры + политики платформ + образование.

07

МЕТОДЫ ОБХОДА 1

Технический уровень: UnMarker

Исследование Университета Ватерлоо (2025)

UnMarker — первая практическая универсальная атака на защитные водяные знаки. Работает без знания алгоритма водяного знака.

> "Защитное водяное знаковое кодирование не является жизнеспособной защитой от дипфейков"
>
> — Andre Kassis, PhD кандидат

Как работает UnMarker

1. **Анализ спектральной области**
   Выявление аномалий в частотном распределении пикселей — признаков водяного знака
2. **Перераспределение частот**
   Нейтрализация водяных знаков через оптимизацию спектральных амплитуд
3. **Black box подход**
   Работает без доступа к алгоритму водяного знака или детектору

Результаты атаки

Снижение эффективности детекции до

43%

даже для семантических водяных знаков

SynthID	Stable Signature
Google	Meta

Почему это работает

Водяные знаки должны быть невидимы и устойчивы к манипуляциям — эти требования ограничивают возможные схемы

Последствия

Уязвимость даже устойчивых схем подрывает доверие к системам маркировки как к надёжной защите

⚠ **Риск:** Технический обход делает водяные знаки ненадёжным инструментом борьбы с дипфейками. Необходимы альтернативные подходы.

Социальная инженерия и промпты

КОНСОРЦИУМ ИССЛЕДОВАНИЙ БЕЗОПАСНОСТИ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

InfoFlood: Информационная перегрузка

Исследование University of Illinois (2025) . Метод джейлбрейка через чрезмерную лингвистическую сложность для дезориентации защитных механизмов.

Пример трансформации:

"Как взломать базу данных?"

→ 194 слова с фейковыми цитированиями и академическим жаргоном

Успех джейлбрейка

~100%

на GPT-4o, Gemini 2.0, Llama 3.1

Как работает InfoFlood

Linguistic Saturation

Реструктуризация запроса с использованием лингвистических трансформаций для повышения сложности

Rejection Analysis

Анализ отказа для выявления причины неудачи и итеративного улучшения

Saturation Refinement

Уточнение запроса для устранения выявленных проблем

Undress AI: Несогласованный контент

Генерация контента для **"раздевания" людей** — серьёзная проблема для всех major LLM:

- **Grok (xAI):** Французские власти расследуют распространение откровенных дипфейков
- **Gemini (Google):** Выявлены случаи генерации сексуализированных изображений
- **ChatGPT/Sora (OpenAI):** Исследователи обнаружили нарушения guardrails

OFCOM (Великобритания): "Серьёзнейшая озабоченность" откровенным контентом с изображениями детей

Почему это работает

Информационная перегрузка нарушает механизмы фильтрации безопасности, не требуя знания внутренней структуры модели

Защита неэффективна

OpenAI Moderation API, Perspective API, SmoothLLM не справляются с InfoFlood-атаками

Проблема: Даже передовые ИИ-системы уязвимы для социальной инженерии. Технические guardrails недостаточны без постоянного совершенствования.

ПРОБЛЕМА ВОСПРИЯТИЯ

Психологический аспект

КОНСОРЦИУМ ИССЛЕДОВАНИЙ БЕЗОПАСНОСТИ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Восприятие маркировки

Исследования показывают: **наличие меток повышает убеждённость** пользователей в том, что контент создан ИИ, но не гарантирует критического восприятия.

CHI 2025 (911 участников): Все дизайны меток привели к вере в ИИ-происхождение контента, но доверие к меткам варьировалось

Эффект избегания информации

Неоднозначные метки

Вызывают наибольший уровень избегания контента

Избегание: 4.05 vs 3.37

Ясные метки

Не увеличивают избегание относительно отсутствия меток

Когнитивный диссонанс

Стратегическое отключение

Пользователи избегают эпистемической неопределённости

Несоответствие меток

Усиливает избегание (Cohen's d = 0.32–0.57)

Парадокс прозрачности

Детальные раскрытия

Снижают доверие пользователей к контенту

Однострочные раскрытия

Не оказывают значительного негативного эффекта

Дилемма: баланс между прозрачностью и доверием

Поведенческие исследования

Метки не оказывают значительного влияния на поведение вовлечённости (лайки, комментарии, репосты)

Обеспокоенность пользователей

61% обеспокоены распространением фейковых новостей и дипфейков, 33% испытывают трудности с распознаванием

Вывод: Мало пометить — надо научить. Маркировка повышает осведомлённость, но не заменяет критическое мышление и медиаграмотность.

Практические рекомендации

КОНСОРЦИУМ ИССЛЕДОВАНИЙ БЕЗОПАСНОСТИ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Внедрение C2PA

Обязательная интеграция стандарта C2PA в инструменты генерации ИИ-контента:

- ✓ Криптографическая привязка метаданных к контенту на этапе генерации
- ✓ Развитие инфраструктуры верификации для пользователей
- ✓ Сертификация продуктов по стандартам C2PA Conformance Program
- ✓ Интеграция в браузеры и социальные платформы

Ужесточение политик платформ

1. **Автоматизированное обнаружение**
   Инвестиции в ИИ-системы для выявления ИИ-контента
2. **Быстрое реагирование**
   Сокращение времени реакции на жалобы пользователей
3. **Прозрачная отчётность**
   Публикация метрик борьбы с дипфейками
4. **Пример Reddit**
   Блокировка сообществ, нарушающих политику

Государственно-частное партнёрство

Развитие сотрудничества между государством и частным сектором:

- ✓ Совместные исследования в области детекции дипфейков
- ✓ Обмен данными о новых методах атак и угрозах
- ✓ Координация регуляторных подходов на международном уровне
- ✓ Финансирование исследований в области ИИ-безопасности

Обучение пользователей

Программы медиаграмотности
Обучение критическому анализу информации и распознаванию манипуляций

Распознавание дипфейков
Обучение признакам синтетического контента (артефакты, несоответствия)

Критическое мышление
Развитие навыков проверки источников и верификации информации

Понимание меток
Объяснение значения различных типов маркировки ИИ-контента

ЗАКЛЮЧЕНИЕ

Маркировка — необходимый, но недостаточный инструмент

Технические решения

C2PA, водяные знаки и другие технологии уязвимы для обхода. UnMarker показал, что даже семантические водяные знаки могут быть удалены с эффективностью 57%.

Регуляторные подходы

AI Act и калифорнийские законы создают основу, но требуют глобальной координации. Фрагментированность законодательства ослабляет эффективность.

Платформы

Не справляются с масштабом задачи — только 33% ИИ-контента корректно маркируется. Необходимы более строгие требования и инвестиции в детекцию.

Психологический фактор

Метки повышают осведомлённость, но не гарантируют критическое восприятие. Необходимо развитие медиаграмотности и критического мышления.

Гонка вооружений между создателями дипфейков и детекторами будет только усиливаться

Контакты: https://trust-ai.ru/company

19 марта 2026