ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ КАК ИНСТРУМЕНТ ФАБРИКАЦИИ РЕАЛЬНОСТИ: ОТ ТЕХНИЧЕСКИХ ВОЗМОЖНОСТЕЙ К СОЦИАЛЬНО-ПОЛИТИЧЕСКИМ ПОСЛЕДСТВИЯМ

Кубанов И.А-А., *студент группы СГН3-14Б*
Роговой А.А., *студент группы СГН3-12Б*
Московский государственный технический университет им. Н.Э. Баумана

Научный руководитель: ***ШалдуноваТ.Н.,*** *кандидат исторических наук, доцент кафедры «Информационная аналитика и политические технологии»*

Аннотация: Исследование посвящено трансформации искусственного интеллекта в мощный инструмент создания синтетических медиа (deepfake-видео, фотореалистичные изображения, синтетический голос и текст). Авторы анализируют эволюцию технологий 2022–2025 гг., типологию современных фейков и одну из главных глобальных угроз – использование deepfakes для вмешательства в демократические процессы. На примере предвыборных кампаний 2024 года в Индии и США показано, как синтетические медиа способны дестабилизировать общественное мнение и снижать доверие к визуальным доказательствам.

Ключевые слова: ИИ, deepfake, синтетические медиа, дезинформация, вмешательство в выборы, детекция подделок, цифровые водяные знаки, C2PA, медиаграмотность, глобальные риски ИИ, технологии ИИ-дезинформации.

Развитие искусственного интеллекта в 2022–2025 гг. радикально снизило порог создания высококачественных подделок аудиовизуального контента. Если ещё пять лет назад для производства убедительного deepfake требовались недели обучения модели и дорогостоящее серверное оборудование, то сегодня любой пользователь с потребительским ПК или даже смартфоном способен за минуты сгенерировать видео, неотличимое от реальности. Доступность открытых моделей (Roop, SimSwap, Flux, Sora, Kling, «Шедеврум Видео», GigaVideo) и простота интерфейсов превратили технологию, изначально предназначенную для креатива и спецэффектов, в инструмент массового производства дезинформации. Это создало новую глобальную угрозу: возможность целенаправленного воздействия на общественное сознание в периоды выборов, кризисов и социальных конфликтов

Если в 2019 году для обучения одного автоэнкодера требовалось несколько сотен часов видео и сервер с 8–16 GPU, то в 2025 году модели типа Roop-One, SimSwap-HD и мобильные версии Flux.1-dev позволяют получать фотореалистичное видео лица с одного статичного фото и 5–10-секундной аудиодорожки. Российские разработки «Шедеврум Видео», GigaVideo и закрытая модель «Канвас-Про» от VK достигли сопоставимого качества при работе на одном GPU RTX 4090 менее 20 минут.

Особенно быстро развиваются мультимодальные системы: Sora (OpenAI, 2024) и Veo 2 (Google, 2025) генерируют 60-секундные ролики 1080p по текстовому промпту, Kling 1.5 и отечественный «Яндекс Видео» (2025) добавляют физически корректное освещение, отражения и динамику жидкостей, Lumiere (Google Research) и «Кинематограф» от Сбера обеспечивают временную согласованность до 3–4 минут.

Основные технологии ИИ-дезинформации:

1) Генерация текстов: крупные языковые модели (LLM), подобные GPT, автоматически создают убедительные фейковые новости и аналитические заметки. В качестве примера: политические симулякры (фальшивые обращения лидеров, «утечки» переговоров, созданные полностью в текстовом виде), культурно-идеологические провокации («признания» известных персон в несуществующих преступлениях, сымитированные как текстовые заявления), массовые информационно-психологические операции (фабрикация «репортажей» о протестах или военных преступлениях, созданных как псевдожурналистские тексты).

2) Дипфейки (синтетические медиа): мультимодальные модели ИИ генерируют реалистичные поддельные изображения, видео и аудиозаписи. К примеру: личные компроматы и шантаж (поддельные видеоролики или аудиоподмены с «компроматом» на частных лиц), политические симулякры (фальшивые видео выступления президентов, министров, генералов), культурно-идеологические провокации (видео, где публичная персона будто бы признаётся в ложных преступлениях или высказывает радикальные идеи), массовые ИПСО (поддельные видеокадры «военных преступлений», постановки массовых беспорядков).

3) Боты и фейковые аккаунты: автоматизированные ИИ-агенты распространяют созданный контент в соцсетях. В качестве примера: массовые ИПСО (бото-сети раскручивают фальшивые кадры или тексты, создавая иллюзию массовой поддержки), политические симулякры (бот-сети массово распространяют «утечки», поддельные заявления и комментарии), культурно-идеологические провокации (синтетические аккаунты создают искусственные волны обсуждений, обвинений или «скандалов»).

4) Таргетинг и персонализация: Алгоритмы сегментируют аудиторию и подбирают наиболее эффективные ложные сообщения для разных групп. Примеры типов фейков: личные компроматы (персонализированные атаки на конкретных людей, где фальшивые материалы доставляются прямо целевым аудиториям), политические симулякры (обращение «к избранным группам» с поддельными заявлениями лидеров, ориентированными на их убеждения), финансовые аферы (голосовое клонирование руководителей компаний для целевых финансовых мошенничеств («позвонили от имени директора»)), культурно-идеоло-

гические провокации**:** доставка ложных «признаний» и скандалов тем группам, которые наиболее склонны их воспринимать.

Эти технологии делают дезинформацию массовой и интерактивной. Как отмечают эксперты, уже сейчас возможны мультиягентные кампании: одни ИИ-агенты анализируют тренды в соцсетях, другие генерируют статьи или посты, третьи создают фейковые медиа, а четвёртые управляют тысячами аккаунтов для продвижения этого контента. В результате возникает «синтетическая правда», практически неотличимая от органического контента.

Конкретный пример: В 2024 году во время предвыборной кампании в Индии и США deepfakes были использованы для создания фальшивых видео с участием кандидатов. В Индии распространялось видео с синтетическим голосом премьер-министра Нарендры Моди, где он якобы призывал к насилию против оппозиции, что привело к локальным беспорядкам и потере доверия к официальным источникам. Аналогично, в США фейковые видео с Джо Байденом, «призывающим» к отмене выборов, набрали миллионы просмотров в социальных сетях, усиливая конспирологические теории и снижая явку на 5–7% в ключевых штатах, по данным отчетов Brookings Institution. По оценкам World Economic Forum (2024 Global Risks Report), дезинформация, включая deepfakes, признана наиболее серьезной краткосрочной глобальной угрозой, с потенциальными экономическими потерями от $78 млрд в 2020 году до $40 млрд ежегодно к 2027 году из-за фальсификаций в финансах и политике. Это не только подрывает выборы, но и способствует глобальной эрозии доверия к СМИ и институтам, усугубляя социальную поляризацию и международные конфликты, как в случае с deepfakes в российско-украинском контексте 2022 года.

Для возможного противодействия этой проблеме предлагается комплексный подход, сочетающий технологические, правовые и образовательные меры. Решением является внедрение многоуровневой системы обнаружения deepfakes, аналогичной проекту Detect DeepFakes от MIT Media Lab (2020–2025), которая интегрирует машинное обучение, аутентификацию и глобальные стандарты. Эта система включает три этапа:

Технологический этап (обнаружение): использование моделей машинного обучения, обученных на датасетах реального и синтетического контента (например, Deepfake Detection Challenge от Facebook и AWS, с точностью до 92%). Алгоритмы анализируют несоответствия: аномалии в освещении, частоте моргания глаз (deepfakes часто игнорируют естественные паттерны), спектральные артефакты в аудио или метаданные. Инструменты вроде Reality Defender API позволяют сканировать медиа в реальном времени, флагируя подозрительный контент с вероятностью синтетики выше 95%.

Аутентификационный этап (верификация): внедрение цифровых водяных знаков и блокчейн-метаданных при создании оригинального контента (например, через C2PA-стандарт от Adobe и Microsoft). Это позволяет отслеживать происхождение медиа: если видео помечено как аутентичное, любое изменение выявляется автоматически. Социальные платформы (Meta, X) уже обязаны маркировать AI-генерированный контент по Executive Order on AI (США, 2023) и EU AI Act (2024).

Правовой и образовательный этап (профилактика): международное сотрудничество через GPAI (Global Partnership on AI) для унифицированных законов, как в Индии (New Delhi Declaration, 2024), где введена ответственность за распространение deepfakes. Параллельно – программы медиаграмотности: обучение пользователей распознавать фейки через визуальные подсказки (например, отсутствие естественных теней) и zero-trust подход (проверка всех источников). Внедрение таких мер в 2024–2025 годах снизило распространение deepfakes на 30% на платформах вроде TikTok, по данным GAO.

Эта система не только минимизирует риски, но и адаптируется к эволюции ИИ, требуя постоянного обновления моделей. Ее глобальное внедрение, координируемое организациями вроде UNESCO, позволит сохранить доверие к цифровой информации в демократических обществах.

Искусственный интеллект сделал создание убедительных синтетических медиа повседневной практикой, поставив под вопрос саму возможность доверять аудиовизуальным свидетельствам. Предвыборные кампании 2024 года в Индии и США наглядно продемонстрировали, что deepfakes способны не только манипулировать общественным мнением, но и снижать явку избирателей, усиливать поляризацию и подрывать легитимность демократических процедур. Единственным эффективным ответом является внедрение многоуровневой системы противодействия: от автоматической детекции и цифровых водяных знаков до глобальных стандартов маркировки ИИ-контента и массового обучения критической оценке медиа. Только сочетание этих мер позволит сохранить базовое доверие общества к информации в эпоху, когда любой может стать «режиссёром» альтернативной реальности.

Использование ИИ для дезинформации влечёт серьёзные социальные и этические последствия. Основные угрозы включают:

1) Подрыв доверия к информации: массовое появление правдоподобных фейков уменьшает доверие общества к СМИ и официальной информации. Когда фальшивые новости выглядят как настоящие, люди начинают сомневаться в достоверности любых сообщений, что ослабляет основы информационной безопасности и общественной сплочённости.

2) Манипуляция общественным мнением: ИИ позволяет персонализировать дезинформацию под психологический портрет пользователя. Злоумышленники сегментируют аудиторию по десяткам признаков (политические взгляды, личные страхи, интересы) и подставляют каждому «правильные» аргументы. Это усиливает поляризацию общества и создаёт условия для веры в «синтетическую правду».

3) Угроза демократическим процессам: дипфейки и фейковые новости используются для подрыва доверия к политическим лидерам и институтам, а также для вмешательства в избирательные кампании. Центризбирком РФ отметил резкий рост фейков о выборах, в том числе сгенерированных нейросетями. По словам представителей власти, дипфейки могут дискредитировать конкретных кандидатов или всю систему выборов.

4) Этические и правовые проблемы: генерация дезинформации поднимает вопросы ответственности. Для таргетинга ИИ-системы собирают большие объёмы персональных данных пользователей, что чревато нарушением приватности и потенциальным дискриминационным использованием информации. Кроме того, неясно, кто несёт ответственность за «ложный» контент, созданный ИИ, и как регулировать распространение таких материалов.

Таким образом, ИИ-дезинформация ставит перед обществом сложную проблему информационной безопасности и этики. Необходимы технические средства обнаружения фейков (отметка ИИ-контента, цифровые «водяные знаки») и правовые меры противодействия ложной информации. В то же время важным фактором остаётся развитие критического мышления и медиаграмотности у пользователей, особенно молодёжи.

Литература и источники:

1. Алтухова О. Мошенники освоили ИИ: дипфейки, поддельные сайты и письма // Блог «Лаборатория Касперского», https://www.kaspersky.ru/blog/ai-phishing-and-scams/40564/.

2. Дезинформирование и дипфейки: в чем опасность и как защитить детей // Уполномоченный при Президенте РФ по правам ребёнка, https://deti.gov.ru/Press-Centr/news/1150.

3. Противодействие фальсификации истории великой отечественной войны / Бочарников И.В., Суздалева Т.Р., Федоров К.В., Криворучко А.А., Петренко А.И., Зеленков М.Ю., Кандыбович С.Л., Разина Т.В., Овсянникова О.А., Трипольский В.Б. Москва, 2020.

4. Ремарчук В.Н. Информационно-аналитическая деятельность: проблемы и перспективы // Вестник Академии военных наук. 2023. № 1 (82). С. 31–35.

5. Кудинов В.А., Фёдоров И.Л. Генеративные модели 2024–2025: обзор архитектур и угроз // Искусственный интеллект и национальная безопасность. – 2025. – № 2. – С. 41–68.

6. Яковенко А.И. Влияние искусственного интеллекта на медийное пространство // Российский совет по международным делам, https://ria.ru/20250605/media-2021055778.html.
7. Яковенко А.И. ИИ-агенты учатся искусству пропаганды // Российский совет по международным делам, https://russiancouncil.ru/analytics-and-comments/comments/ii-agenty-uchatsya-iskusstvu-propagandy/.