Л.Н. Гумилев атындағы Еуразия ұлттық университетінің ХАБАРШЫСЫ.
ISSN (Print) 2616-7174. ISSN (Online) 2663-2500

БАҚ ЖӘНЕ ҚОҒАМ / СМИ И ОБЩЕСТВО /MEDIA AND SOCIETY

МРНТИ 19.01.07
Научная статья

https://doi.org/10.32523/2616-7174-2026-155-2-79-100

Дипфейки в медиаэкосистеме: синтетическая информация и ее влияние на цифровую аудиторию

Ш. И. Калиаждарова[1*], С. В. Ашенова[2], Ж. Д. Сейтжанова[3]

[1,2,3]*Международный университет информационных технологий, Алматы, Казахстан*

(E-mail: [1]s.kaliazhdarova@iitu.edu.kz, [2] s.ashenova@iitu.edu.kz, [3] zh.seitzhanova@iitu.edu.kz)

Аннотация. В статье исследуется феномен deepfake, который рассмотрен не как изолированная технологическая новация, а как структурный элемент современного медиапространства, встроенный в логику цифрового производства и эмоционального вовлечения аудитории. Цель исследования заключается в анализе влияния AI-сгенерированного медиаконтента на процессы восприятия информации и формирования доверия к СМИ, а также в разработке типологического подхода к анализу дипфейков. Методологическую основу работы составляют концепции медиадоверия, цифровой дезинформации и синтетических медиа, а также элементы эмпирического анализа, применяемые для классификации различных форм дипфейк-контента. В результате исследования показано, что дипфейки представляют собой системную проблему медиасреды, усиливающую риски манипуляции, пропаганды и дезинформации и способствующую эрозии доверия цифровой аудитории к медиаконтенту. Обоснована необходимость классификации дипфейков не только по техническим характеристикам, но прежде всего по степени воздействия и семантическому назначению, включая манипулятивные, пропагандистские, дезинформационные и иронические формы. Делается вывод о том, что существующие журналистские и редакционные практики противодействия синтетическим медиа, включая инструменты верификации, обладают ограниченной эффективностью без изменения отношения аудитории к потребляемой и распространяемой информации. Устойчивость информационной среды в цифровом обществе требует перераспределения ответственности между медиапрофессионалами и пользователями, а также развития превентивной журналистики и аналитических форматов.

Ключевые слова: дипфейки, синтетические медиа, доверие к СМИ, цифровая аудитория, дезинформация, манипуляция.

Поступила: 30.03.2026; Одобрена: 29.05.2026; Доступна онлайн: 30.06.2026

***автор для корреспонденции**

Введение

Современные исследования определяют deepfake как синтетически сгенерированный или модифицированный медиаконтент (видео, изображение, аудио), созданный с использованием методов глубокого обучения и генеративных моделей (например, GAN) и имитирующий реальных людей с высокой степенью реализма. Такие технологии позволяют создавать визуально правдоподобные сцены, которые трудно отличить от аутентичных материалов, что увеличивает риск их злоупотребления в медийном пространстве. Возникает вопрос – готово ли общество к тому, чтобы самостоятельно определять критерии доверия к общественной информации без помощи медиаспециалистов? Какие технологии могут противостоять сгенерированной информации? Какие цели преследует такая информация, или, вернее, те, кто ее создает? И как профессиональная журналистика может научить аудиторию ориентироваться в новостном потоке, содержащем в себе методы и технологии deepfake.

Эрозия общественного доверия к традиционным СМИ является сегодня одним из ключевых вызовов, обсуждаемых в научном дискурсе. Высокореалистичные deepfake видео и изображения увеличивают скептицизм аудитории, поскольку зрители не уверены в подлинности визуальных свидетельств, что уменьшает доверие к журналистике как институции. Сегодня deepfakes рассматриваются как потенциальный инструмент для распространения ложной информации, особенно в контексте политических событий, конфликтов и социально-значимой информации (Ahmed, 2020). Это создает дополнительные сложности для журналистов, которые обязаны верифицировать материалы и противостоять манипуляции общественным мнением.

Актуальность настоящего исследования обусловлена необходимостью адаптации журналистской деятельности к условиям цифровой медиасреды, характеризующейся стремительным распространением синтетического медиаконтента, создаваемого с использованием технологий искусственного интеллекта. Масштабирование технологий deepfake формирует новые риски для достоверности информации и общественного доверия к средствам массовой коммуникации, поскольку даже профессиональные участники медиапроцесса нередко сталкиваются с трудностями при идентификации фальсифицированных аудиовизуальных материалов. В данной связи особую значимость приобретает внедрение технологических инструментов детектирования синтетического контента в редакционные процессы.

Целью данного исследования является анализ существующих подходов к распознаванию deepfake-контента в цифровых мультимедийных материалах, создаваемых с применением генеративных нейронных сетей, их классификация, а также оценка применимости соответствующих методов и инструментов в профессиональной журналистской практике. Объектом исследования являются цифровые мультимедийные материалы, создаваемые с использованием алгоритмов генеративных нейронных сетей.

Предметом исследования являются дипфейк-технологии как форма медиаконтента, классифицируемая по функциональному назначению на дезинформационные, манипулятивные, пропагандистские и сатирические практики, а также особенности их воздействия на общественное восприятие информации.

В рамках данного исследования предполагается комплексный анализ теоретических и прикладных подходов к определению и типологизации deepfake-контента,

представленных в современной научной литературе. Особое внимание уделяется рискам, которые ИИ-сгенерированные мультимедийные материалы представляют для общественного доверия, процессов формирования общественного мнения и механизмов медиаверификации. На основе этого осуществляется систематизация существующих методов и инструментов распознавания deepfake-контента, а также оценка их эффективности и ограничений в условиях реального медиапроизводства и редакционной практики с анализом функциональных типы дипфейков – дезинформационные, манипулятивные, пропагандистские и сатирические, с точки зрения целей их создания и механизмов воздействия на общественное мнение. Дополнительно рассматриваются особенности интерпретации AI-сгенерированного контента аудиторией и профессиональные риски, с которыми сталкиваются журналисты при работе с подобными материалами, что позволяет сформулировать практические рекомендации по интеграции технологий распознавания дипфейков в современную журналистскую деятельность, включая практики превентивной и проверочной журналистики.

Обзор литературы

В ноябре 2025 года президентом РК Касым-Жомартом Токаевым был официально утвержден закон «Об искусственном интеллекте», который демонстрирует наличие четко сформулированного законодательства по AI и создает основу, на которую могут опираться исследования deepfake-контента. Сам факт регулирования, включая требования к обязательной маркировке синтетического контента и ограничения на его использование, предоставляет ученым законодательные ориентиры для анализа эффективности технических методов обнаружения deepfake и оценки правовых рисков при их применении в медиасреде. Следует уточнить, что Закон включает положения о прозрачности, безопасности и ответственности при использовании AI-систем, а также классификацию AI по уровню риска. Это стимулирует исследовательские проекты, которые объединяют правовые, технические и журналистские подходы, например, анализ эффективности технических инструментов в контексте соответствия требованиям законодательства или оценку влияния нормативных мер на редакционные процессы. Социальная специфика остается недостаточно проработанной. Наше общество не является идеальным в плане потребления цифровой информации, как, впрочем, и любой социум сегодня, что приводит к появлению новых цифровых вызовов.

Некоторые исследования поднимают вопросы этики применения deepfake технологий в медиа: использование синтетического контента в новостях или репортажах без должной маркировки нарушает принципы прозрачности и ответственности журналистики. Однако вопрос применения AI в журналистике намного глубже. Значительная часть исследований интерпретирует распространение deepfake в контексте кризиса доверия к медиа. D.Sotirova (Sotirova, 2025) подчеркивает, что технологии искусственного интеллекта усиливают существующий скептицизм аудитории по отношению к новостным источникам, поскольку визуальный контент, ранее воспринимавшийся как наиболее надежный, утрачивает статус объективного свидетельства. По мнению автора, deepfake не создают принципиально новых форм манипуляции, но радикально увеличивают их масштаб и правдоподобие, что осложняет поддержание журналистской легитимности.

Схожие выводы представлены в работе E.Lundberg (Lundberg, 2024) посвященной анализу потенциального влияния deepfake на новостные медиа. Автор отмечает

двойственную природу данной технологии: при наличии легитимных сценариев применения (реконструкция событий, визуализация исторического контекста) deepfake одновременно способствуют формированию феномена «тотального сомнения», при котором аудитория склонна подвергать сомнению подлинность даже проверенных материалов. Это обстоятельство оказывает косвенное, но устойчивое негативное влияние на журналистику как социальный институт.

Отдельное направление исследований связано с проблемой дезинформации, манипулятивными технологиями и возможностями AI как медиа инструмента. С. К. Pandey и соавторы (Pandey, 2025) рассматривают deepfake не только как угрозу, но и как инструмент, применимый в творчестве и медиапрактиках как один из наиболее эффективных инструментов современной манипуляции общественным мнением, при этом авторы обсуждают возможности и потенциал AI-технологий наряду с рисками.

Дополняя данные подходы, A. Rauchfleisch и соавторы (Rauchfleisch, 2025) акцентируют внимание на дискурсивном измерении проблемы. Авторы показывают, что сам термин «deepfake» в публичном и медийном дискурсе способствует усилению восприятия рисков и может усугублять недоверие к визуальной журналистике. В этой связи подчеркивается необходимость коммуникативных стратегий со стороны СМИ, направленных на разъяснение принципов работы синтетических медиа и границ их применения.

В целом анализ научной литературы свидетельствует о том, что влияние технологий deepfake на журналистику носит комплексный и многоуровневый характер. Исследователи сходятся во мнении, что ключевыми последствиями являются эрозия доверия к новостному контенту, рост сложности верификационных процедур и усиление этических дилемм журналистской практики. Однако теоретические исследования, проводимые на стыке научных дисциплин, использующие междисциплинарный подход и формирующие концептуальные основы понимания этих феноменов в медиапространстве, остаются недостаточно разработанными. С одной стороны, технические обзоры подробно описывают алгоритмические методы генерации и обнаружения deepfake, анализируя существующие подходы и методы оценки эффективности детекторов, однако фокусируются преимущественно на инженерных аспектах, а не на концептуальных медиа-теоретических моделях влияния и противодействия этим технологиям в сообщениях СМИ. На этом фоне применительно к интересам журналистики более значимо выглядят практико-ориентированные исследования. Результаты, полученные H. Nawaz с соавторами свидетельствуют о том, что даже осведомленность о существовании deepfake технологий может приводить к снижению доверия к новостям, независимо от фактического столкновения аудитории с фальсифицированным контентом (Nawaz., Hashmi, 2024). Авторы делают вывод о системном характере угрозы, поскольку снижение доверия затрагивает журналистику в целом, а не отдельные медиаорганизации.

Одним из наиболее цитируемых эмпирических исследований последних лет является работа C. Vaccari и A. Chadwick, посвященная воздействию политических deepfake-видео на восприятие новостного контента. На основе экспериментального дизайна с репрезентативной выборкой британской аудитории авторы показали, что даже в отсутствие прямого обмана deepfake-контент способствует росту эпистемической неопределенности и снижению доверия к новостным источникам. Тем самым

deepfake рассматриваются не только как инструмент дезинформации, но и как фактор системной эрозии доверия к медиасреде (Vaccari., Chadwick, 2020). Ряд исследований фокусируется на сравнительном анализе различных форматов дипфейков. S.Ahmed & соавторы на практике продемонстрировали, что видео-дипфейки воспринимаются пользователями как более убедительные по сравнению с аудио- и текстовыми фейками и чаще вызывают намерение дальнейшего распространения. Статья, в которой представлены экспериментальные данные опросов более восьми тысяч человек из восьми стран, анализирует предикторы склонности делиться дипфейками в социальных сетях (например, восприятие точности, страх упустить что-то важное, саморегуляция, когнитивные способности и использование соцсетей). Исследование показало, что интенсивность распространения дипфейков связана с восприятием их правдоподобности, низкой саморегуляцией и низкой когнитивной способностью (Ahmed, 2020). Эти результаты подчеркивают особую уязвимость визуальных медиа и указывают на повышенные риски для новостных редакций, работающих с видеоконтентом.

Важное направление эмпирических исследований связано с изучением способности пользователей распознавать deepfake-контент. Международное сравнительное исследование J. Frank и соавторов, проведенное в США, Германии и Китае, показало, что точность распознавания искусственно сгенерированных медиа у респондентов лишь незначительно превышает уровень случайного угадывания. Отсутствие существенных межстрановых различий свидетельствует о структурном характере данной проблемы и ограниченности индивидуальных когнитивных стратегий в условиях высокореалистичного синтетического контента (Frank., Herbert, 2023).

Анализ научной литературы показывает, что эмпирические исследования deepfake-контента представлены широким спектром методологических подходов и формируют значительный массив данных о восприятии, воздействии и рисках, связанных с синтетическими медиа. Вместе с тем эти исследования в большинстве случаев остаются фрагментированными и недостаточно интегрированными в целостные теоретические модели, что снижает их объяснительный и прикладной потенциал. Отсутствие унифицированных концептуальных рамок затрудняет разработку эффективных стратегий реагирования, поскольку различные типы дипфейков предполагают различную логику воздействия и, соответственно, требуют дифференцированных правил противодействия. В результате возрастает потребность в структурно обоснованных моделях, способных заложить концептуальные основы для анализа deepfake-контента и выстраивания редакционных практик, сравнительных медийных стратегий и механизмов противодействия на уровне медиасистем для разработки концептуальных моделей медиапрактик.

Методология

Методологическая основа исследования носит комплексный и междисциплинарный характер, что обусловлено технологической и социально-коммуникативной природой феномена deepfake. В работе используется совокупность теоретических и эмпирических научных методов, направленных на выявление их ключевых технических и визуальных признаков, а также на их классификацию в зависимости от целей использования в медиапространстве.

Применимые в этом контексте научные методы позволяют проанализировать природу и механизмы создания дипфейков, выявить их ключевые технические и

визуальные признаки, а также классифицировать виды дипфейков в зависимости от целей использования. Метод контент-анализа позволяет определить систематическое выявление повторяющихся визуальных и технических характеристик дипфейков и позволяет количественно и качественно описать набор их типичных признаков. Мы можем выделить следующие типы: дезинформация, сатира, пропаганда, манипуляция общественным мнением. Выделение типов дипфейков осуществлялось на основе анализа коммуникативных целей и контекста использования синтетического контента. Это обеспечило теоретическую базу для дальнейшего анализа, так как вышеуказанный метод используется для анализа видеоматериалов и аудиовизуального контента с целью фиксации несоответствий мимики и артикуляции; нарушений синхронизации речи и изображения; аномалий освещения, текстур кожи, движения глаз и микромимики; артефактов компрессии и генерации изображения.

Для выявления смысловых и выразительных особенностей визуального контента использован метод визуально-семиотического анализа, который позволяет выявить как визуальные элементы дипфейков имитируют достоверность и «эффект реальности», усиливая при этом эмоциональное воздействие и маскируя тем самым искусственное происхождение контента. Этот метод особенно эффективен при различении сатиры и намеренной манипуляции.

Научную классификацию дипфейков по целям использования позволил определить типологический метод, благодаря которому была произведена группировка дипфейков на основе коммуникативного намерения, контекста распространения, предполагаемого эффекта воздействия на аудиторию, что, в свою очередь, предоставило возможность выделить основные их типы и обратиться к методу критериального анализа для формирования четких оснований, способных распределить кейсы по типам. Необходимость сопоставления дипфейков разных типов между собой определил обращение к методу сравнительного анализа, который позволил выявить как общие тенденции, так и специфические особенности каждого типа в контексте журналистской практики. Ключевым для интерпретации рисков, связанных с использованием дипфейков в журналистике, стал контекстуальный анализ, применяемый для учета социальных, политических и медиа условий создания и распространения дипфейков, включая специфику платформ, целевые аудитории и сопутствующие текстовые и метаданные.

Обобщение полученных результатов осуществлено с использованием метода междисциплинарного синтеза, позволяющего интегрировать технические, визуальные и коммуникативные аспекты анализа в единую аналитическую модель оценки рисков и перспектив использования дипфейков в современной журналистике.

Применение данного методологического подхода позволяет не только выявить риски, связанные с распространением дипфейков в журналистике, но и определить потенциальные направления адаптации профессиональных стандартов и практик в условиях развития синтетических медиа.

Результаты и обсуждение

В современной медиаэкосистеме дипфейки перестали быть маргинальным технологическим феноменом и превратились в важную и неотъемлемую часть цифрового мира, оказывающую прямое влияние на информационные процессы, общественное сознание и характер современных коммуникационных конфликтов.

Сначала deepfake использовался в сфере развлечений: для шуток, пародий, фан-видео. Однако очень быстро он начал применяться в более опасных контекстах – от создания фальшивых новостей и дискредитации публичных лиц до фишинговых атак и распространения дезинформации. Текущие отчеты ООН и ЕС ясно связывают deepfake с угрозами общественному доверию, вмешательству в выборы и необходимости усиления мер обнаружения и противодействия синтетическому контенту во всех плоскостях цифрового сообщества.

Современная медиасреда характеризуется доминированием социальных сетей и мессенджеров, которые в значительной степени взяли на себя функции основных каналов распространения общественно значимой информации. В этих условиях границы между институционально подтвержденными сообщениями, частными интерпретациями и намеренно искаженными данными становятся все менее различимыми для аудитории. Поскольку большинство традиционных средств массовой информации – включая телевизионные каналы, печатные издания и информационные агентства – активно используют цифровые платформы для дистрибуции контента, практика получения новостей через социальные сети стала для пользователей преобладающей. В результате доверие все чаще формируется не на основе проверки источника и фактической достоверности информации, а исходя из самого канала ее распространения. Присутствие сообщения в онлайн-пространстве нередко воспринимается как достаточное условие его правдоподобия, что снижает мотивацию аудитории обращаться к первоисточникам и официальным ресурсам СМИ.

В рамках обозначенных процессов особую роль начинают играть технологии создания дипфейков, усиливающие рискогенный потенциал медиасреды. Синтетически сгенерированные аудиовизуальные материалы, основанные на алгоритмах глубокого обучения, не только встраивается в информационные потоки, но и трансформирует сами механизмы доверия и интерпретации медиареальности. Их способность имитировать визуальные и вербальные характеристики реальных лиц способствует размыванию границ между достоверной и искаженной информацией, что ускоряет переход информационных рисков в категорию информационных угроз и оказывает прямое влияние на формирование цифровой идентичности и общественного мнения. Данный подход развивает положения классических медиатеорий, в частности, концепции симуляции и гиперреальности, а также теории медиатизации, где логика познания соотносится с положениями классических медиатеорий, прежде всего с концепцией симуляции и гиперреальности Ж. Бодрийяра (Baudrillard, 1999). В своих работах «Симулякры и симуляция» и «Прозрачность зла» он утверждал, что общество живет не в реальности как таковой, а в пространстве знаков, которые больше не отсылают к оригиналу. Симулякр не искажает реальность, он замещает ее, создавая гиперреальность, где различие между подлинным и сконструированным утрачивает значение. Применяя его интерпретацию к современным способам передачи информации, можно утверждать, что многие социальные медиа-источники производят знаки, не отсылающие к реальности, а замещающие ее, формируя автономное пространство смыслов. Дипфейки в этом контексте могут рассматриваться как симулякры третьего порядка, то есть медиаобъекты, не искажающие действительность как таковую, а функционирующие независимо от нее, что делает различие между подлинным и сконструированным вторичным. С этой позиции дипфейки могут рассматриваться как завершенная форма

симуляции, в рамках которой медиареальность перестает отражать действительность и начинает функционировать автономно. Тем самым дипфейки не просто усиливают информационные риски, а знаменуют переход к состоянию гиперреальности, где категории подлинности, авторства и факта утрачивают свою регулятивную функцию. Этот тезис находит свое развитие и в современных исследованиях AI, где подчеркивается, что алгоритмически созданный контент оценивается аудиторией не по критерию истинности, а по степени правдоподобия и эмоциональной убедительности. Участники современного цифрового пространства воспринимают видеодипфейки как более достоверные и склонны делиться ими, так как у аудитории существует склонность к субъективному восприятию реалистичности, а не проверяемости фактов (Ahmed, 2020). Это открывает большое поле для игры с дезинформацией, как одного из описываемых нами типов deepfake.

Уже в начале полномасштабного конфликта между Россией и Украиной в марте 2022 г. в цифровом медиапространстве был зафиксирован синтетически сгенерированный видеоматериал, в котором использовался образ президента Украины Владимира Зеленского, это было видео, в котором президент Украины якобы призывает вооруженные силы сложить оружие и сдаться. Этот материал получил большое распространение через социальные сети и даже появился в бегущей строке телеканала Ukraine 24, прежде чем был опровергнут и удален с платформ Meta и других ресурсов. Эксперты установили, что видео было сгенерировано с помощью технологий искусственного интеллекта. Хотя современные инструменты deepfake создают настолько убедительные видео, что визуально отличить их от оригинала становится все труднее, однако на техническом уровне все еще можно выявить следы синтетического вмешательства с помощью следующих методов, которые и были применены в данном случае:

Микроартефакты изображения – легкое «мыление», артефакты компрессии, неправильные тени или размытие в области лица;

Неестественные движения глаз – у многих deepfake-моделей глаза моргают либо слишком редко, либо одновременно, с одинаковой продолжительностью;

Несовпадение мимики с речью – расхождение между движением губ и звуком;

Неправильные отражения и освещение – например, свет может падать на лицо с другого угла, чем на остальные части тела;

Искажения при повороте головы или сильной мимике – многие модели плохо обрабатывают резкие эмоции, гримасы или быстрые повороты.

В казахстанском сегменте информационного пространства в октябре 2025 года были зафиксированы видеоматериалы, в которых, предположительно, участвовали представители Министерства иностранных дел и Министерства обороны Республики Казахстан. В указанных материалах содержались утверждения о причастности Казахстана к атаке беспилотных летательных аппаратов на нефтеперерабатывающий завод в городе Орск. Анализом занялся Центр по борьбе с дезинформацией, который выявил признаки генерации искусственного интеллекта, после чего официальные ведомства Казахстана опровергли эту информацию. Дипфейки распространялись через социальные сети и мессенджеры, создавая ложное впечатление о международных событиях и вовлечении Казахстана в конфликт, что могло привести к искажению восприятия ситуации казахстанской и зарубежной аудиторией. Центр определил такие технические признаки, как: «Несовпадение мимики и артикуляции с голосом –

характерное «залипание» лица и неестественные тени при повороте головы. Искаженная структура звука (отсутствие естественных шумов помещения, сжатие аудиосигнала, типичное для синтетического голоса). Нарушение синхронизации между губами и речью при просмотре в замедленном режиме. Метаданные видеофайлов, распространяемых в Telegram и X (Twitter), не содержат признаков исходной съемки: отсутствуют данные о модели устройства и времени записи» (Ondato, 2025).

По оценке Центра по борьбе с дезинформацией, подобные информационные вбросы направлены на расшатывание общественных настроений в Казахстане и России, конструирование нарратива о вовлеченности Казахстана в конфликтные процессы, а также эрозию доверия к официальной коммуникации и дипломатическим институтам (AntiFake Center, 2025).

Исходя из устойчивого положения о том, что дезинформация определяется распространением ложного контента, вводящего в заблуждение, при которой цель может быть как четко определена, так и иметь скрытую этимологию (Lada.kz, 2025) при этом не обязательно с политической подоплекой, но с возможностями значимых международных последствий, эти примеры мы можем отнести к дезинформационному типу deepfake, так как фальшивые видео, приписывающие официальные заявления лицам, имеющим политическую, в некоторых случаях социально значимую власть, демонстрируют риск быстрого распространения ложной информации, которая может стать причиной межгосударственных недоразумений.

Это, несомненно, связано с рисками, воздействующими на качество медиа, поскольку воздействие на аудиторию посредством конструирования реальности основано на существующих событиях, способных формировать информационную повестку дня. Аналогичным образом манипулятивные подходы могут применяться для формирования общественных представлений о мире и его этических нормах. Это достигается за счет изменения семантической значимости информации в виртуализированном информационном пространстве, создаваемом с использованием нейросетевых технологий и способном вводить аудиторию в заблуждение.

Следует отметить, что дезинформация, применяемая как манипулятивная технология в медиапространстве, не обязательно должна преследовать ярко окрашенную политическую или идеологическую цель, но манипуляция в ней присутствует практически всегда. Вопрос в том, каким образом ожидаемый результат должен воздействовать на конечного потребителя. Если в политической дезинформации, рассмотренной выше, целевые установки носят явный международно-дестабилизационный подтекст, то в примере с deepfake с Папой Римским в пуховике, вышедшем в 2023 году, визуальная манипуляция имела под собой культурную составляющую с отсутствием обоснования как такового и с основным эффектом в виде размывания границ между реальным и синтетическим. Манипуляция присутствовала, но не имела прямого влияния на общественное мнение по конкретному вопросу, оставляя за собой психологический эффект воздействия на аудиторию. Поэтому и этот кейс стоит отнести к дипфейк-дезинформации в отличие, например, от фейкового видео с Нэнси Пелоси, первой женщиной-спикером Палаты представителей Конгресса США, которое было опубликовано в сети Facebook в 2019 году. Оно не просто вводило в заблуждение, а подрывало доверие к конкретному политическому актору и использовалось в политическом контексте для его дискредитации, что является признаком манипуляции общественным мнением.

Дезинформация как таковая здесь вторична, так она стала не целью, а способом повлиять на электорат. Мы можем сделать вывод о том, что учет условий и среды распространения дипфейков, включая медиаплатформы, целевые аудитории и сопутствующие текстовые и метаданные, характерные для контекстуального анализа, в таких случаях становится ключевым методом для различения дезинформации и манипуляции общественным мнением.

В то же время техногенные риски, выходя на новый уровень цифровых технологий, вносят изменения и в устойчивые парадигмы определяемых понятий манипуляции и дезинформации. В 2024 году на просторах интернет появилось аудиосообщение, сгенерированное с применением технологий искусственного интеллекта и имитирующее голос президента США Джо Байдена. Оно представляет собой показательный пример одновременного сочетания дезинформации и манипуляции общественным мнением.

Содержательно запись распространяла ложную информацию, приписывая главе государства позицию, которой он фактически не занимал, в то время как функционально она была направлена на воздействие на электоральное поведение зарегистрированных сторонников Демократической партии в ходе праймериз в штате Нью-Гэмпшир. По заявлению генерального прокурора Джона Формеллы данный аудиодипфейк был создан техасской компанией Life Corporation, принадлежащей Уолтеру Монку, связанному с бизнесами в сфере политических автоматизированных обзвонов и телефонных опросов. Высокая степень акустического сходства с реальным голосом президента, включая использование характерных речевых оборотов, усиливала убедительность сообщения и, как следствие, повышала его манипулятивный потенциал.

Это яркий пример появления гибридных форм синтетической дезинформации, сочетающих в себе признаки как дезинформации, так и манипуляции общественным мнением, что стало возможным благодаря развитию технологий генеративного искусственного интеллекта. В отличие от традиционных форм ложного контента, такие материалы не ограничиваются распространением недостоверных сведений, но одновременно эксплуатируют доверие к аудиовизуальным медиа, усиливая эмоциональное воздействие и искажая процессы общественного восприятия реальности. Гибридный характер синтетической дезинформации проявляется в том, что ложное сообщение, будучи формально идентифицируемым как дезинформация, функционально выполняет манипулятивную задачу, то есть направленное влияние на установки, поведение и политические решения аудитории. Использование дипфейк-технологий позволяет воспроизводить узнаваемые визуальные и аудиальные маркеры публичных фигур, что существенно повышает убедительность контента и снижает порог критического восприятия со стороны пользователей.

В результате синтетическая дезинформация функционирует не только как инструмент информационного воздействия, но и как фактор структурного изменения медиасреды, трансформируя способы производства, распространения и потребления новостей. Таким образом, гибридные формы синтетической дезинформации следует рассматривать не как частный технологический феномен, а как системный медиариск, требующий комплексного анализа на пересечении медиаисследований, политической коммуникации и исследований цифровой безопасности.

Кейс с аудиодипфейком, имитирующим голос президента Джо Байдена, вынудил ведущие редакции, включая The New York Times и Washington Post, отказаться от

воспроизведения самой записи и сосредоточиться на анализе механизмов манипуляции, источников распространения и институциональных последствий. Параллельно регуляторы и правоохранительные органы инициировали обсуждение необходимости правового регулирования синтетического политического контента, что косвенно указывает на признание угрозы на государственном уровне. В европейском контексте рост числа визуальных и аудиодипфейков стал одним из факторов принятия Digital Services Act и AI Act, (AntiFake Center, 2025), предусматривающих обязательную маркировку синтетического контента и усиление ответственности платформ. Европейские СМИ стали внедрять редакционные протоколы, ограничивающие повторное распространение дипфейков даже в разоблачающем контексте, что отражает переход от реактивной к превентивной модели журналистики.

В условиях цифровой медиасреды это может рассматриваться как форма проактивной коммуникации, ориентированной на раннее выявление и интерпретацию общественно значимых рисков. Понятие эксклюзивности, сенсационности, необходимости первыми сообщить значимую информацию, присущее журналистике раннего цифрового расцвета, сегодня должно трансформироваться в модальность работы с общественным мнением, имеющего как долгосрочную, так и ситуативную перспективу. Нельзя не обращать на это внимание, так как современная массовая аудитория в свете своих потребительских привычек и новых требований к информации в большей степени подвержена трендовым характеристикам с высокой степенью интерактивности, фрагментированности и участия в производстве контента. Принимая во внимание понимание публичной сферы как пространства рационально-критического обсуждения, что является базовой основой исследовательских подходов к формированию общественного мнения (Sultanbayeva., Tolegen, 2024), анализ пользовательской активности в цифровых платформах, включая комментарии, социальные сети и пользовательские обращения, позволит журналистике фиксировать зарождающиеся проблемы до их перехода в острую фазу. В этом контексте аудитория выступает не только объектом информирования, но и источником эмпирических данных, расширяющих аналитический потенциал медиа. Таким образом, превентивная журналистика в цифровой среде смещает акцент с реактивного освещения событий на упреждающее объяснение потенциальных последствий, что усиливает общественную функцию медиа и способствует снижению социальных и медиарисков.

В регионах с повышенной конфликтностью медиариски проявляются особенно остро. Использование синтетических видео и аудиозаписей в контексте вооруженных конфликтов, вынуждает международные редакции создавать специализированные подразделения по верификации визуального контента. При этом сами журналисты признают, что даже успешное разоблачение фейков не всегда восстанавливает доверие аудитории, что усиливает эффект так называемого «информационного истощения» (Vaccari., Chadwick, 2020).

Примером комплексного дипфейка может служить информаионная ситуация вокруг конфликта Израиль-Газа, 2023–2024 гг., когда в социальных сетях широко циркулировали AI-генерированные изображения, которые демонстрировали эмоционально тяжелые сцены разрушения. Эти изображения, созданные с помощью искусственного интеллекта, получили вирусное распространение в Facebook, Twitter и других платформах, часто сопровождаясь ложными утверждениями о том, что они являются реальными кадрами с места боевых действий. Тщательный анализ экспертов доказал характерные визуальные

аномалии на изображениях, такие как неестественные пропорции, искажения пальцев и сияние глаз, все это указывало на синтетическое происхождение материалов, которые распространялись с целью вызвать эмоциональную реакцию и усилить восприятие конфликта как чрезвычайно жестокого и трагичного. Эти элементы делали их инструментом эмоциональной манипуляции и в то же время являлись дезинформацией. Однако принцип дезинформации в данном случае уступал место манипуляции общественным мнением как целенаправленного влияния, а не просто введения в заблуждения, что и явилось в итоге основным фактором для определения типологии deepfake.

В данном кейсе преследовалась четкая политическая цель оправдания военных действий, демонизации противника и эмоционального давления на аудиторию, когда фокус восприятия смещался с информационного на эмоциональный, что типично для пропаганды и манипулятивных техник, когда важен уже не факт, а ощущение правоты, так как сторонники нарратива продолжают в него верить и распространять в доступных социальных сетях и мессенджерах. Именно этот момент становится знаковым для определения манипуляции и пропаганды, но приносит большие неудобства для работы журналистов, существенно ограничивая возможности позитивного интерактива со своей аудиторией. Манипулятивные нарративы, усиленные или полностью сгенерированные средствами искусственного интеллекта, распространяются обычно с такой высокой скоростью, что адаптация под эмоциональные и ценностные установки различных аудиторных групп происходит практически мгновенно, и это затрудняет их своевременную идентификацию и деконструкцию. Для журналистов становится проблематичным фиксировать момент перехода от интерпретации к манипуляции, поскольку AI-контент часто сохраняет формальные признаки достоверности, включая визуальную убедительность. Это снижает потенциал превентивной журналистики как механизма общественного информирования и требует пересмотра профессиональных стандартов и аналитических подходов в работе с AI в отношении к эмоционально насыщенному и алгоритмически оптимизированному распространению манипуляции.

В Казахстане с 18 января 2026 г. вступили в силу нормы, обязывающие обязательно маркировать материалы, созданные с помощью искусственного интеллекта, включая дипфейки, как синтетические, с указанием соответствующих визуальных или машинно-читаемых предупреждений (Orda.kz, 2026).

Эти правила распространяются не только на СМИ, но и на все интернет-ресурсы, и предусматривают административные и уголовные санкции за нарушение маркировки или распространение ложного контента вплоть до трех лет лишения свободы в случае значительного вреда или угрозы общественному порядку. На национальном уровне Казахстан также развивает механизмы мониторинга цифрового пространства через государственную систему «Кибернадзор» для выявления и анализа противоправного содержания, включая синтетические материалы, с последующим направлением сведений компетентным органам для реагирования.

В медиа РК манипуляция может быть определена как устойчивая лексема, имеющая границы спецификаций. Например, манипуляция через «опросы общественного мнения», где присутствует селективный выбор лексики, допустимая «повестка дня» или предлагается заведомо безальтернативный вывод. Такие практики формируют ограниченное поле смыслов, когда сложно определить техники информационных

манипуляций, что в свою очередь требует развития аналитических инструментов, ориентированных на дискурсивный и нарративный анализ. При этом в медиапространстве Казахстана присутствовали сатирические дипфейки, маркированные как «сатирические новости», например, Агентство сатирических новостей Qaznews24, которое прекратило свою деятельность в 2025 году, но имело успех в социальных сетях, при этом зачастую пользователи делились информацией, предлагаемой каналом, как правдивой, не обращая внимания на маркировку, что в некотором смысле послужило косвенной причиной закрытия паблика. В то же время международный опыт демонстрирует открытость социума такому типу дипфейков как художественному комментарию или деструктивной критике. Примером могут служить проекты с использованием дипфейк-технологий, но в контексте демонстрации возможностей легитимного сатирического использования при сохранении прозрачности и контекстуальной ясности.

В США популярен сатирический веб-проект Sassy Justice, в котором эти технологии применяются для создания вымышленных ситуаций с участием узнаваемых публичных фигур, что дает критическое осмысление политической коммуникации и медийных конвенций, при этом пародийный характер контента не маскируется и является частью авторского замысла. Проект создан Trey Parker, Matt Stone (авторы популярного South Park) и Peter Serafinowicz, где с помощью дипфейка лица известных политических и медийных фигур помещаются в абсурдный контекст жизни маленького городка и расследований выдуманного репортера. Это не попытка ввести в заблуждение, а комедийная социальная сатира на медиаманипуляции и политическое общение. Zizi&Me/The Zizi Show: AI-перформанс и кабаре. Проект британского художника Jake Elwes сочетает дипфейк с перформансом в стиле drag-cabaret, где искусственный аватар выступает вместе с артистом, сатирически осмысливая представления о теле, идентичности и роли AI в культуре. Это выставочное художественное использование дипфейка, близкое к сатире, где технология функционирует не как средство имитации реальности, а как художественный прием, подчеркивающий условность цифровых образов и медиапредставлений. В рамках художественных медиапроизведений и институциональных кампаний создавались дипфейки с участием политиков для социальных комментариев или благотворительных целей. Президент Франции Эммануэль Макрон сам опубликовал серию подделок с его образом в развлекательных сценах, где демонстрировались танцы и пародийные клипы, чтобы привлечь внимание к Саммиту по искусственному интеллекту. Они носили юмористический, не политически манипулятивный характер и были отмечены пользователями как игра с медийным образом. Таким образом, сатирические дипфейки не только принципиально отличаются от дезинформации и манипуляции, так как используют технологии для комментария или критики, а не для убеждения в ложном факте, и должны быть четко маркированы как пародия или художественный продукт, но и вполне осязаемо несут положительно-конструктивный заряд для развития превентивной журналистики.

В то же время повсеместные усилия по правовому регулированию, внедрению маркировки и развитию кадровых и технических редакционных стандартов отражают в первую очередь признание угрозы дипфейков как международного медиариска. Эти меры направлены не только на пресечение распространения дезинформации, но и на создание устойчивых механизмов защиты доверия аудитории к журналистике в эпоху генеративного AI. При этом законодательные меры демонстрируют фрагментарность

и асимметричность. Если одни государства делают ставку на регулирование платформ и маркировку AI-контента, другие используют риторику борьбы с дезинформацией для расширения контроля над медиасредой. Это создает дополнительный риск для журналистики, поскольку инструменты противодействия синтетической дезинформации могут быть использованы как для защиты общественного интереса, так и для ограничения свободы слова.

Редакционные политики крупных СМИ выстраиваются таким образом, чтобы снижать риск распространения манипулятивных форм визуальной или аудиальной дезинформации и поддерживать доверие аудитории. Это отражается в типичных редакционных стандартах по борьбе с дезинформацией и модифицированными медиа согласно практике фактчекинга в крупных новостных организациях.

В Казахстане эта динамика нашла отражение в развитии профессионального фактчекинга: проект Factcheck.kz, основанный в 2017 году как первый профессиональный фактчекинговый ресурс в Центральной Азии и действующий на основе принципов Международной сети фактчекинга (IFCN), в свое время был включен в глобальную Программу сторонней проверки фактов корпорации Meta для работы с вирусной информацией на платформах Facebook, Instagram и Threads, что позволяло снижать распространение ложного контента и маркировать его для аудитории. Работа ресурса включает в себя анализ и проверку сомнительной информации, особенно той, которая активно циркулирует в социальных сетях, как это было с сообщениями о псевдоизменениях в правилах дорожного движения в 2025 году.

Одновременно государственные усилия по противодействию дезинформации получили институциональную форму через создание Центра по борьбе с дезинформацией при Центральной службе коммуникаций, который активно идентифицирует и опровергает ложные сообщения, например о массовых отключениях электроэнергии, подготовке беспорядков или ложных нормативных изменениях.

В рамках исследования был проведен опрос среди молодых людей в возрасте от 18 до 30 лет, согласно которому было выявлено, что значительная доля потребителей затрудняется отличить проверенную информацию от фейков, а многие не прибегают к верификации новостей. Большинство из них отметили, что доверяют новостям, которые узнают из социальных сетей и если новость предоставляется официальными СМИ, они не будут ее перепроверять, но, если информацию прокомментирует или опровергнет селебрити, они, скорее всего, поверят ему, а не журналистам. На первый план выходит психоэмоциональная айдентика: люди охотно позволяют лидерам мнений принимать за них важные решения и направлять их по различным вопросам. Это происходит потому, что цифровизация социализирует аудиторию как единую массу. Хотя такие подходы не являются новым открытием – ранее они рассматривались в контексте пропаганды и маркетинга – именно в эпоху цифрового контента, с беспрецедентным масштабом социализации через интернет, современная аудитория вступила в парадоксальную фазу. Возможность определять себя как личность со свободой выбора и мнения во Всемирной паутине растворилась в массе мнений большинства трендов и цифровых модных веяний по широкому кругу тем.

Это, на наш взгляд, стало одним из важных источников проблемы фейковых новостей, которая основана не только на возможностях современных информационных технологий, но и на психоэмоциональной составляющей цифровой личности.

В продолжение эксперимента той же группе молодых людей было предложено оценить важность самоанализа информации, предлагаемой в интернете. Основной посыл был следующим: «Готовы ли вы нести ответственность за информацию, которую самостоятельно получаете из различных источников, и продвигать ее дальше?» Большинство респондентов решили, что им нужен помощник, например, в виде чат-бота, который проверял бы информацию, даже из официальных СМИ. Доверие к «лидерам мнений» снизилось почти на 25%, когда встал вопрос об их собственной ответственности. В то же время число тех, кто хотел бы использовать цифровые инструменты для проверки информации, увеличилось. В ходе эксперимента участникам было предложено использование Telegram-бота, самостоятельно разработанного студентами кафедры журналистики. В течение трех дней участники отправляли боту как реальные изображения из новостных источников, так и заранее подготовленные фотографии в формате дипфейк, полученные с помощью генераторов нейронных сетей.

Большинство респондентов выразили заинтересованность в расширении его функциональности за счет добавления анализа видео, текстовых новостей и введения кнопки «Отправить на проверку» в браузере. Некоторые участники отметили, что бот может быть полезен в первую очередь в повседневной практике редакций. Были также те, кто был готов полностью полагаться на результаты проверки информации без дальнейшего анализа, а также те, кто не был заинтересован в эксперименте.

Кроме того, студентам и преподавателям, специализирующимся на системах информационной безопасности, а также преподавателям, специализирующимся на информационной безопасности в сфере массовых коммуникаций, были заданы вопросы: «Действительно ли безопасно и абсолютно необходимо внедрение искусственного интеллекта, создающего информацию различных типов, для аудитории?» и «Доверяете ли вы искусственному интеллекту и насколько?». Ответы оказались очень интересными. 91% респондентов готовы постоянно использовать искусственный интеллект, но 22% полностью ему доверяют. Они считают, что искусственный интеллект способен совершать серьезные ошибки – 70%, и в то же время человек способен полностью контролировать генерацию его дальнейшего развития – 56%.

Преподаватели, представляющие различные подходы к использованию и изучению искусственного интеллекта, представители технических и социальных областей единодушно пришли к выводу – взаимодействие с искусственным интеллектом и его возможностями, включая генерацию контента, в значительной степени является философским вопросом, фундаментом, заложенным в детстве. Те же системы информационной безопасности направлены на воспитание людей, которые будут противостоять противоправным действиям, совершаемым с использованием искусственного интеллекта, поскольку его использование тесно связано с цифровым обществом.

Таким образом, проблема взаимодействия человека с искусственным интеллектом выходит за пределы сугубо технологического или регуляторного измерения и приобретает ценностно-нормативный характер. Эффективное противодействие негативным характеристикам ИИ, включая генерацию манипулятивного контента, во многом определяется уровнем сформированности критического мышления и этических установок, закладываемых на ранних этапах социализации. В этом контексте алгоритмы информационной безопасности и медиаобразования следует рассматривать не только

как инструменты технического контроля, но и как элементы долгосрочной стратегии формирования ответственного цифрового поведения в условиях развития цифрового общества. Совокупность рисков, возникающих в результате распространения deepfake-контента, свидетельствует о необходимости системного изучения данного формата как структурного элемента современной медиасреды, обладающего различными типологиями и неоднородными механизмами воздействия на аудиторию.

Заключение

Если обратиться к технологиям обнаружения deepfake-контента, то можно выделить несколько ключевых направлений, в которых разрабатываются соответствующие инструменты. Наиболее важными для современной цифровой медиасреды являются гибридные платформы, которые комбинируют несколько подходов, например, AI FactCheckers, анализирующая как изображение, так и текстовую часть поста. Однако все эти платформы и технологии deepfake основаны на алгоритмах машинного обучения, прежде всего на генеративно-состязательных сетях (GAN, Generative Adversarial Networks).

При этом ключевая проблема заключается в том, что в условиях стремительного развития технологий синтетических медиа феномен deepfake выходит за рамки исключительно технического обнаружения и требует междисциплинарного осмысления в рамках журналистики, медиакоммуникаций и системы информационной безопасности. Несмотря на расширение инструментов автоматизированного распознавания deepfake, их эффективность имеет тенденцию расти в соответствии со скоростью распространения контента и адаптивностью манипулятивных практик. В этой связи ключевым фактором устойчивости информационной среды становится способность аудитории критически интерпретировать медиасообщения и различать формы использования сгенерированного контента – от сатиры и художественного высказывания до дезинформации, манипуляции и пропаганды.

Исследование «deepfakes as narratives» подтверждает, что люди во многих странах не способны уверенно отличить синтетические медиа от реальных, при этом склонность оценивать искусственный контент как «человеческий» сохраняется (Lada.kz, 2025). Проведенные авторами опросы, кроме того, показали, что пользователи видеоконтента часто переоценивают собственную способность распознавать дипфейки, что делает их более восприимчивыми к убедительным, но ложным визуальным сообщениям. В то же время эмоциональное погружение и восприятие нарратива усиливают положительную оценку дипфейков, независимо от знания об их искусственном происхождении. Таким образом, дипфейки следует рассматривать не как изолированное технологическое явление, а как структурный элемент современной медиасреды, в которой классические теоретические положения о симуляции, медиавоздействии и социальной конструкции реальности получают новое развитие в контексте искусственного интеллекта. Их распространение ускоряет переход информационных рисков в плоскость информационных угроз и оказывает прямое влияние на процессы формирования цифровой идентичности, общественного сознания и механизмов социального взаимодействия.

Для журналистики это означает смещение акцента с исключительно фактчекинговых функций к образовательной и превентивной роли, направленной на формирование медиаграмотности и осознанного восприятия цифровых образов. Следовательно,

развитие навыков распознавания манипулятивных и пропагандистских дипфейков у аудитории становится неотъемлемой частью современной информационной экологии и важным условием обеспечения информационной безопасности в цифровом обществе.

Разнообразиеструктуритипологийдипфейков–отсатирическихиэкспериментальных до манипулятивных и пропагандистских – предполагает неодинаковые механизмы воздействия на аудиторию и, соответственно, различный уровень общественных рисков. В этом контексте превентивная журналистика сталкивается с принципиальной сложностью: отсутствие своевременной классификации и интерпретации AI-сгенерированного контента ограничивает возможность раннего предупреждения аудитории о потенциальных манипуляциях. В этой связи особое значение приобретают практики медиаграмотности и обучения распознаванию дипфейков, интегрированные в социальные сети и цифровые платформы, где распространение манипулятивного контента происходит наиболее активно и где пользователи чаще становятся невольными участниками его тиражирования. Поддержка аудитории через образовательные, интерфейсные и коммуникационные инструменты распознавания синтетических медиа представляется необходимым условием снижения деструктивного воздействия дипфейков и формирования более устойчивой и ответственной цифровой медиасреды, которая сегодня сталкивается с необходимостью мгновенной проверки информации в условиях информационного шума. Чем быстрее распространяется недостоверный или потенциально опасный контент, тем выше риск того, что он станет «новой истиной» в общественном сознании. Влияние такого контента не ограничивается ареалом цифровой личности: он влияет на политические взгляды, общественное мнение и на повестку дня социума.

Журналистика, выступая как институт общественной ответственности, оказывается ключевым актором в выстраивании аналитических рамок, позволяющих не только фиксировать факт существования дипфейка, но и типологизировать его, объяснять его структуру, функциональное назначение и возможные последствия, тем самым снижая его манипулятивное влияние на восприятие информации в цифровом пространстве.

Вклад авторов:

Калиаждарова Ш.И. – предложила идею исследования и написала основной текст статьи.

Ашенова С.В. – собрала данные, провела статистический анализ и интерпретировала результаты. Международный университет информационных технологий.

Сейтжанова Ж.Д. – подготовила обзор литературы и отредактировала статью.

Благодарность, конфликт интересов

Представленная статья не финансировалась из внешних источников и выполнена в форме аналитического материала. Авторы выражают благодарность студентам кафедры медиакоммуникаций и истории Казахстана, преподавателям АО МУИТ за участие в социологическом опросе. Авторы заявляют об отсутствии конфликта интересов.

Список литературы:

Ahmed, S., Ng, S. W. T. and Bee, A. W. T. (2020) Understanding the role of fear of missing out and deficient self-regulation in sharing of deepfakes on social media: Evidence from eight countries. Frontiers in Psychology, 14, Article 1127507. DOI: https://doi.org/10.3389/fpsyg.2023.1127507.

Baudrillard, J. (1999) Simulacra and simulation (S. F. Glaser, Trans.). University of Michigan Press (Original work published 1981, Éditions Galilée).

Дипфейк: от имени казахстанских чиновников распространяются ложные заявления о «причастности Казахстана к атаке на Орск». [Онлайн] Қолжетімді: https://t.me/AntiFakeCenter/56 (қаралған күні: 25.01.2026).

Frank, J., Herbert, F., Ricker, J., Schönherr, L., Eisenhofer, T., Fischer, A., Dürmuth, M. and Holz, T. (2023) A representative study on human detection of artificially generated media across countries [Preprint]. Cornell University: arXiv. [Онлайн] Қолжетімді: https://arxiv.org/abs/2312.05976 (қаралған күні: 16.07.2026).

Frank, J., Herbert, F., Ricker, J., Schönherr, L., Eisenhofer, T., Fischer, A., Dürmuth, M. and Holz, T. (2024) A representative study on human detection of artificially generated media across countries. In: Proceedings of the 45th IEEE Symposium on Security and Privacy. San Francisco, California, USA: IEEE, pp. 55-73. DOI: https://doi.org/10.1109/SP54263.2024.00159.

Habermas, J. (1962) The structural transformation of the public sphere: An inquiry into a category of bourgeois society (T. Burger, Trans.). MIT Press (Original work published 1962).

Hashmi, M., Shahzad, K., Lin, W., Tsao, Y. and Wang, H.-M. (2024) Unmasking illusions: Understanding human perception of audiovisual deepfakes [Preprint]. Cornell University: arXiv. [Онлайн] Қолжетімді: https://arxiv.org/abs/2405.04097 (қаралған күні: 16.07.2026).

Hwang, Y. and Lee, J. (2025) The potential effects of deepfakes on news media and entertainment. AI & SOCIETY, 40, pp. 2159-2170. DOI: https://doi.org/10.1007/s00146-024-02072-1.

Lundberg, E. and Mozelius, P. (2024) The potential effects of deepfakes on news media and entertainment. AI & SOCIETY, 40(4), pp. 2159–2170. DOI: https://doi.org/10.1007/s00146-024-02072-1.

Deepfake Laws: Global Overview and Emerging Regulations. [Онлайн] Қолжетімді: https://ondato.com/blog/deepfake-laws/ (қаралған күні: 9.01.2026).

Pandey, A. K., Minhas, D., Goyal, E., Bhowmik, M., Khajanchi, Y. and Dhawan, A. (2025) Deepfake art: Threat or innovation in digital creativity. ShodhKosh: Journal of Visual and Performing Arts, 6(1S), pp. 32–41. DOI: https://doi.org/10.29121/shodhkosh.v6.i1s.2025.6619.

Rauchfleisch, A., Vogler, D. and de Seta, G. (2025) Deepfakes or Synthetic Media? The Effect of Euphemisms for Labeling Technology on Risk and Benefit Perceptions. Social Media + Society, 11(3), pp. 1-13. DOI: https://doi.org/10.1177/20563051251350975.

Shu, K., Wang, S., Lee, D. and Liu, H. (2020) Disinformation, misinformation, and fake news in social media: Emerging research challenges and opportunities. Cham, Switzerland: Springer International Publishing.

Sotirova, D. (2025) AI and deepfake technology in times of crisis of trust in media: Is it really a grave threat to journalism? Medialog, 17, pp. 43-53. DOI: https://doi.org/10.60060/MLg.2025.17.43-53.

Г.С. Султанбаева and Б.З. Толеген (2024) Распространение фейковой и некорректной

информации в социальных сетях. Вестник КазНУ. Серия журналистика, 72(2), pp. 681-403. DOI: https://doi.org/10.26577/HJ.2024.v72.i2.3.

Vaccari, C. and Chadwick, A. (2020) Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social Media + Society, 6(1), Article 2056305120903408, pp. 1–13. DOI: https://doi.org/10.1177/2056305120903408.

В Казахстане будут штрафовать за ИИ-контент без маркировки. [Онлайн] Қолжетімді: https://orda.kz/v-kazahstane-budut-shtrafovat-za-ii-kontent-bez-markirovki-410898/ (қаралған күні: 18.01.2026).

В Сети появилось видео, которое ставит Казахстан под подозрение. [Онлайн] Қолжетімді: https://www.lada.kz/kazakhstan-news/144147-v-seti-poiavilos-video-kotoroe-stavit-kazakhstan-pod-podozrenie.html (қаралған күні: 25.01.2026).

Ш.И. Калиаждарова[1], С.В. Ашенова[2], Ж.Д. Сейтжанова[3]

[1,2,3]Халықаралық ақпараттық технологиялар университеті, Алматы, Казахстан

Медиаэкожүйесіндегі Дипфейк: синтетикалық ақпарат және оның цифрлық аудиторияға әсері

Аңдатпа. Мақалада дипфейк феномені жеке технологиялық инновация ретінде емес, цифрлық медиақызметтің өндірістік логикасына және аудиторияны эмоциялық тұрғыдан тарту механизмдеріне кіріктірілген қазіргі заманғы медиакеңістіктің құрылымдық құрамдас бөлігі ретінде қарастырылады. Зерттеудің мақсаты — жасанды интеллект негізінде генерацияланған медиаконтенттің ақпаратты қабылдау үдерістеріне және бұқаралық ақпарат құралдарына деген сенімнің қалыптасуына ықпалын талдау, сондай-ақ дипфейк-контентті зерттеудің типологиялық тәсілін ұсыну. Зерттеудің әдіснамалық негізін медиасенім, цифрлық дезинформация және синтетикалық медиа теориялары, сондай-ақ дипфейктердің әртүрлі формаларын жүйелеуге бағытталған эмпирикалық талдау элементтері құрайды. Зерттеу нәтижелері дипфейктердің медиакеңістіктегі жүйелік сипатқа ие құбылыс екенін, олардың манипуляция, үгіт-насихат және дезинформация тәуекелдерін арттырып, цифрлық аудиторияның медиаконтентке деген сенім деңгейінің төмендеуіне ықпал ететінін көрсетеді. Дипфейктерді жіктеудің техникалық параметрлермен шектелмей, олардың ықпал ету деңгейі мен семантикалық бағытталуына негізделген типологиясын әзірлеу қажеттілігі негізделеді. Атап айтқанда, манипулятивтік, насихаттық, дезинформациялық және ироniялық формалар айқындалады. Мақалада синтетикалық медиамен күресуге бағытталған қолданыстағы журналистік және редакциялық тәжірибелердің, соның ішінде верификация құралдарының, аудиторияның ақпаратты тұтыну және тарату мәдениетін трансформациялаусыз шектеулі тиімділікке ие екені туралы тұжырым жасалады. Цифрлық қоғам жағдайында ақпараттық ортаның тұрақтылығын қамтамасыз ету медиамамандар мен пайдаланушылар арасындағы жауапкершілікті қайта бөлуге, сондай-ақ превентивті журналистика мен аналитикалық форматтарды дамытуға тәуелді екені негізделеді.

Түйін сөздер: дипфейк, синтетикалық медиа, бұқаралық ақпарат құралдарына сенім, цифрлық аудитория, дезинформация, манипуляция, үгіт-насихат, медиакеңістік

Sh. I. Kaliazhdarova[1], S.V. Ashenova[1], Seytzhanova Zh.D.[3]
[1,2,3]International Information Technologies University, Almaty, Kazakhstan

Deepfakes in the Media Ecosystem: Synthetic Information and Its Impact on Digital Audiences

Abstract. This article examines the deepfake phenomenon, not as an isolated technological innovation, but as a structural element of the modern media landscape, integrated into the logic of digital production and emotional audience engagement. The aim of the study is to analyze the impact of AI-generated media content on information perception and the formation of trust in media, as well as to develop a typological approach to analyzing deepfakes. The methodological basis of the work is the concepts of media trust, digital disinformation, and synthetic media, as well as elements of empirical analysis used to classify various forms of deepfake content. The study demonstrates that deepfakes represent a systemic problem in the media environment, increasing the risks of manipulation, propaganda, and disinformation, and contributing to the erosion of digital audience trust in media content. The need to classify deepfakes not only by technical characteristics, but primarily by their degree of impact and semantic purpose, including manipulative, propaganda, disinformation, and ironic forms, is substantiated. The study concludes that existing journalistic and editorial practices for countering synthetic media, including verification tools, are limited in effectiveness without changing audience attitudes toward the information they consume and disseminate. The sustainability of the information environment in a digital society requires a redistribution of responsibility between media professionals and users, as well as the development of preventive journalism and analytical formats.

Keywords: deepfakes, synthetic media, media trust, digital audience; disinformation, manipulation, propaganda.

References

Ahmed, S., Ng, S. W. T. and Bee, A. W. T. (2020) Understanding the role of fear of missing out and deficient self-regulation in sharing of deepfakes on social media: Evidence from eight countries. Frontiers in Psychology, 14, Article 1127507. DOI: https://doi.org/10.3389/fpsyg.2023.1127507.

Baudrillard, J. (1999) Simulacra and simulation (S. F. Glaser, Trans.). University of Michigan Press (Original work published 1981, Éditions Galilée).

Dipfeyk: ot imeni kazakhstanskikh chinovnikov rasprostranyayutsya lozhnye zayavleniya o «prichastnosti Kazakhstana k atake na Orsk» [Deepfake: false statements about "Kazakhstan's involvement in the attack on Orsk" are being spread on behalf of Kazakhstani officials] (2025). [Online] Available at: https://t.me/AntiFakeCenter/56 (Accessed: 25 January 2026). [In Russian]

Frank, J., Herbert, F., Ricker, J., Schönherr, L., Eisenhofer, T., Fischer, A., Dürmuth, M. and Holz, T. (2023) A representative study on human detection of artificially generated media across countries [Preprint]. Cornell University: arXiv. Available at: https://arxiv.org/abs/2312.05976 (Accessed: 16 July 2026).

Frank, J., Herbert, F., Ricker, J., Schönherr, L., Eisenhofer, T., Fischer, A., Dürmuth, M. and Holz,

T. (2024) A representative study on human detection of artificially generated media across countries. In: Proceedings of the 45th IEEE Symposium on Security and Privacy. San Francisco, California, USA: IEEE, pp. 55-73. DOI: https://doi.org/10.1109/SP54263.2024.00159.

Habermas, J. (1962) The structural transformation of the public sphere: An inquiry into a category of bourgeois society (T. Burger, Trans.). MIT Press (Original work published 1962).

Hashmi, M., Shahzad, K., Lin, W., Tsao, Y. and Wang, H.-M. (2024) Unmasking illusions: Understanding human perception of audiovisual deepfakes [Preprint]. Cornell University: arXiv. Available at: https://arxiv.org/abs/2405.04097 (Accessed: 16 July 2026).

Hwang, Y. and Lee, J. (2025) The potential effects of deepfakes on news media and entertainment. AI & SOCIETY, 40, pp. 2159-2170. DOI: https://doi.org/10.1007/s00146-024-02072-1.

Lundberg, E. and Mozelius, P. (2024) The potential effects of deepfakes on news media and entertainment. AI & SOCIETY, 40(4), pp. 2159–2170. DOI: https://doi.org/10.1007/s00146-024-02072-1.

Ondato (2025) Deepfake Laws: Global Overview and Emerging Regulations. [Online] Available at: https://ondato.com/blog/deepfake-laws/ (Accessed: 9 January 2026).

Pandey, A. K., Minhas, D., Goyal, E., Bhowmik, M., Khajanchi, Y. and Dhawan, A. (2025) Deepfake art: Threat or innovation in digital creativity. ShodhKosh: Journal of Visual and Performing Arts, 6(1S), pp. 32–41. DOI: https://doi.org/10.29121/shodhkosh.v6.i1s.2025.6619.

Rauchfleisch, A., Vogler, D. and de Seta, G. (2025) Deepfakes or Synthetic Media? The Effect of Euphemisms for Labeling Technology on Risk and Benefit Perceptions. Social Media + Society, 11(3), pp. 1-13. DOI: https://doi.org/10.1177/20563051251350975.

Shu, K., Wang, S., Lee, D. and Liu, H. (2020) Disinformation, misinformation, and fake news in social media: Emerging research challenges and opportunities. Cham, Switzerland: Springer International Publishing.

Sotirova, D. (2025) AI and deepfake technology in times of crisis of trust in media: Is it really a grave threat to journalism? Medialog, 17, pp. 43-53. DOI: https://doi.org/10.60060/MLg.2025.17.43-53.

Sultanbayeva, G.S. and Tolegen, B.Z. (2024) Rasprostranenie feykovoy i nekorrektnoy informatsii v sotsial'nykh setyakh [Spread of fake and incorrect information in social networks]. Vestnik KazNU. Seriya zhurnalistika, 72(2), pp. 681-403. DOI: https://doi.org/10.26577/HJ.2024.v72.i2.3. [In Russian]

Vaccari, C. and Chadwick, A. (2020) Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social Media + Society, 6(1), Article 2056305120903408, pp. 1–13. DOI: https://doi.org/10.1177/2056305120903408.

V Kazakhstane budut shtrafovat' za II-kontent bez markirovki [In Kazakhstan, they will fine for AI content without labeling] (2026). [Online] Available at: https://orda.kz/v-kazahstane-budut-shtrafovat-za-ii-kontent-bez-markirovki-410898/ (Accessed: 18 January 2026). [In Russian]

V Seti poyavilos' video, kotoroe stavit Kazakhstan pod podozrenie [A video has appeared on the Web that puts Kazakhstan under suspicion] (2025). [Online] Available at: https://www.lada.kz/kazakhstan-news/144147-v-seti-poiavilos-video-kotoroe-stavit-kazakhstan-pod-podozrenie.html (Accessed: 25 January 2026). [In Russian]

Сведения об авторах:

Калиаждарова Ш. И. – автор для корреспонденции, доктор PhD, ассоциированный профессор, Международный университет информационных технологий, ул Манаса, 34/1, 050040, Алматы, Казахстан.

Ашенова С.В. – кандидат политических наук, ассоциированный профессор, Международный университет информационных технологий, ул Манаса, 34/1, 050040, Алматы, Казахстан.

Сейтжанова Ж.Д. – кандидат филологических наук, ассоциированный профессор, Международный университет информационных технологий, ул Манаса, 34/1, 050040, Алматы, Казахстан.

Авторлар туралы мәлімет:

Қалиаждарова Ш. И. – хат-хабар авторы, PhD, қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университеті, Манаса көшесі 34/1, 050040, Алматы, Қазақстан.

Ашенова С.В. – саяси ғылымдар кандидаты, қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университеті, Манаса көшесі 34/1, 050040, Алматы, Қазақстан.

Сейтжанова Ж.Д. – филология ғылымдарының кандидаты, қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университеті, Манаса көшесі 34/1, 050040, Алматы, Қазақстан.

Information about the authors:

Kaliazhdarova Sh. I. – Corresponding author, PhD, Associate Professor, International University of Information Technology, Manas Street 34/1, 050040, Almaty, Kazakhstan.

Ashenova S.V. – Candidate of Political Sciences, Associate Professor, International University of Information Technology, Manas Street 34/1, 050040, Almaty, Kazakhstan.

Seytzhanova Zh.D. – Candidate of Philological Sciences, Associate Professor, International University of Information Technology, Manas Street 34/1, 050040, Almaty, Kazakhstan.