IPIE

Международная группа экспертов по информационной среде

Реагирование на дезинформацию, создаваемую генеративным ИИ

Результаты метаанализа научных данных

Резюме для лиц, принимающих решения
2026.2

Номер DOI: 10.61452/QXAF2136

A. Herasimenka, S. Valenzuela, S. Boulianne, F. Esser, L. M. Given, S. Lewandowsky, E. M. Navarro-Lopez, P. N. Howard

Реагирование на дезинформацию, создаваемую генеративным ИИ

Результаты метаанализа научных данных

Резюме для лиц, принимающих

Формат цитирования:
International Panel on the Information Environment [A. Herasimenka, S. Valenzuela, S. Boulianne, F. Esser, L. M. Given, S. Lewandowsky, E. M. Navarro-López, P. N. Howard (eds.)], "Responding to Generative AI Misinformation: Results from a Meta-Analysis of Scientific Evidence" [«Реагирование на дезинформацию, создаваемую генеративным ИИ: результаты метаанализа научных данных»], Zurich, Switzerland: IPIE, 2026. Summary for Policymakers [Резюме для лиц, принимающих решения], SFP2026.2, doi: 10.61452/NLZH2351.

КРАТКИЙ ОБЗОР

Генеративный искусственный интеллект (GenAI) в настоящее время способен быстро создавать большие объемы вводящих в заблуждение текстов, изображений, аудио- и видеоматериалов с помощью легкодоступных инструментов. Это имеет принципиальное значение для лиц, принимающих решения, поскольку ложный или вводящий в заблуждение контент поддаётся целенаправленной адаптации под конкретные аудитории. Подобный контент способен быстро распространяться и имитировать подлинную коммуникацию, что затрудняет для людей оценку достоверности информации.

В данном Резюме для лиц, принимающих решения, в сжатом виде представлены основные выводы Обобщающего доклада Международной группы экспертов по информационной среде (IPIE) (SR2026.2) о последствиях воздействия дезинформации, созданной с помощью GenAI, и мерах, которые с наибольшей вероятностью снизят её влияние.

Данная оценка опирается на крупномасштабный метаанализ экспериментальных научных данных. Она основана на 60 оценках эффектов по результатам рандомизированных контролируемых исследований из 24 рецензируемых публикаций с участием 33 801 человека, опубликованных в период с 2018 по 2025 год.

В докладе сформулированы четыре основных вывода:

1. Дезинформация в текстовом формате, созданная GenAI, в настоящее время представляет более высокие риски с точки зрения убеждающего воздействия, чем визуальная дезинформация.
2. Текущая доказательная база не охватывает большую часть мира. Исследования сосредоточены в англоязычных странах и странах с высоким уровнем дохода, что оставляет значительные пробелы в наших знаниях.
3. Наиболее стабильно эффективным вмешательством является превентивное предоставление пользователям корректирующей информации, чтобы они могли самостоятельно оценивать точность и достоверность.
4. Маркировка эффективна при условии ее единообразного применения. Маркировка контента, как правило, снижает воспринимаемую достоверность, однако ее воздействие сильно варьируется в зависимости от того, как компании создают и применяют метки.

Выводы для государственной политики очевидны. Лицам, принимающим решения, необходимо сделать противодействие дезинформации в текстовом

формате, создаваемой GenAI, первоочередным приоритетом и содействовать применению превентивной и корректирующей информации в качестве основополагающей стратегии. К маркировке следует относиться как к мере, требующей тщательной разработки и тестирования. Мы должны расширить независимые исследования далеко за пределы англоязычной среды и стран с высоким уровнем дохода.

Доступ исследователей к данным платформ и моделей имеет решающее значение для углубления общественного понимания проблемы дезинформации, созданной GenAI, и для тестирования того, какие защитные механизмы эффективны на практике.

ВВЕДЕНИЕ

Системы генеративного искусственного интеллекта (GenAI) способны создавать текст, изображения, аудио- и видеоматериалы, которые людям порой трудно отличить от контента, созданного человеком. Контент, сгенерированный ИИ, получил широкое распространение, фигурируя на выборах по всему миру, часто из неизвестных или злонамеренных источников, и все больше влияя на общественное восприятие в таких сферах, как здравоохранение, конфликты и рынки, что влечет за собой значительные последствия в реальном мире [1]. Эти системы включают в себя большие языковые модели, генерирующие текст, инструменты синтеза изображений и звука, а также системы создания видео. Они также могут создавать вводящий в заблуждение или вредоносный контент, такой как фейковые новости, слухи, пропаганда или дипфейки. В данном докладе эти явления в совокупности именуются дезинформацией.

В данном Резюме для лиц, принимающих решения, освещаются основные выводы метаанализа экспериментальных данных о последствиях воздействия на людей дезинформации, созданной GenAI [2]. В нем также оцениваются две меры противодействия, подкрепленные достаточным объемом экспериментальных данных для их обсуждения при разработке политики: корректирующая информация и маркировка контента. В докладе особый акцент делается на рандомизированных контролируемых исследованиях, поскольку эти работы обеспечивают более надежные сопоставления результатов различных экспериментов и позволяют проводить более систематическую оценку средних эффектов.

Полученные данные свидетельствуют о быстро развивающейся информационной среде. В докладе отмечается, что реакция общественности на текстовую и визуальную дезинформацию в настоящее время демонстрирует расхождения. В нем также делается вывод о том, что доказательная база остается ограниченной по своему географическому охвату и методологии. В то же время в нем определены две практические меры реагирования, способные снизить воспринимаемую достоверность дезинформации, сгенерированной GenAI, хотя одна из них демонстрирует более стабильную эффективность, чем другая.

Для формирования этой доказательной базы в рамках обзора был проведен поиск в двух крупных библиографических базах данных, Scopus и Web of Science, учтены рекомендации экспертов, объединены результаты, удалены дубликаты и проанализирован итоговый массив из 6 952 публикаций. После проверки на соответствие критериям отбора и полнотекстового кодирования 87 публикаций были признаны удовлетворяющими требованиям для детального обзора. Дополнительные проверки на наличие

риска систематической ошибки, скрининг и проверки протоколов в соответствии с рекомендациями Кокрейновского сотрудничества привели к выделению подгруппы из 24 наиболее релевантных исследований для включения в метаанализ. В основном Обобщающем докладе применяется трехуровневый метаанализ со случайными эффектами и оценивается риск систематической ошибки с использованием Пересмотренного инструмента оценки риска систематической ошибки в рандомизированных исследованиях. Подробная информация о методологии [2] представлена в полном тексте Обобщающего доклада.

РЕЗУЛЬТАТ 1. ВВОДЯЩИЙ В ЗАБЛУЖДЕНИЕ ТЕКСТ GENAI ПРЕДСТАВЛЯЕТ НАИБОЛЬШУЮ УГРОЗУ

ОСНОВНОЙ ВЫВОД

Дезинформация в текстовом формате, созданная генеративным ИИ, в настоящее время представляет более высокие риски с точки зрения убеждающего воздействия, чем визуальная дезинформация.

В докладе выявляется четкое различие в восприятии достоверности текстового и визуального контента, сгенерированного GenAI. Более поздние исследования показывают, что ложный текст, сгенерированный GenAI, часто воспринимается как более точный, достоверный или правдоподобный, чем подлинная информация. В то же время исследования ложной визуальной информации, созданной GenAI, такой как дипфейки, показывают, что она часто воспринимается как менее достоверная, чем подлинная информация. В исследованиях визуального контента, проведенных после 2021 года, объединенная оценка указывает на умеренное снижение воспринимаемой достоверности.

Это не означает, что визуальная дезинформация безопасна или не имеет значения. Дипфейки по-прежнему представляют риск, и в докладе отмечается, что реакция общественности может вновь измениться по мере совершенствования визуальных систем. Однако текущие данные свидетельствуют о том, что текстовая дезинформация заслуживает более безотлагательного внимания при разработке политики. Это имеет важное значение, поскольку системы, ориентированные на работу с текстом, могут быть недорогими, масштабируемыми, легко настраиваемыми и все чаще интегрируются в поисковые системы, мессенджеры и диалоговые интерфейсы. В докладе также указывается, что результаты генерации диалогового ИИ могут быть значительно более убедительными, чем статические сообщения, сгенерированные ИИ, что позволяет предположить, что текущие оценки могут быть консервативными.

Общая картина полученных результатов достаточно убедительна для того, чтобы сформулировать четкий посыл для лиц, принимающих решения: наиболее непосредственный риск убеждающего воздействия в настоящее время исходит от ложного или вводящего в заблуждение текста, а не только от визуальных дипфейков.

РЕЗУЛЬТАТ 2. ГЛОБАЛЬНЫЕ ПРОБЕЛЫ В ДОКАЗАТЕЛЬНОЙ БАЗЕ

В докладе четко подчеркивается, что большинство изученных публикаций сосредоточено на англоязычных странах и странах с высоким уровнем

дохода, таких как Австрия, Германия, Нидерланды, Южная Корея, Великобритания и США. Текущая доказательная база ограничена полнотекстовыми статьями на английском языке, а пять проанализированных публикаций на других языках не соответствовали критериям отбора. Следовательно, текущее понимание проблемы дезинформации, созданной GenAI, сосредоточено на небольшом числе стран, языков и медийных сред.

ОСНОВНОЙ ВЫВОД

Текущая доказательная база оставляет большую часть мира: страны и языки — за рамками рассмотрения.

Это является существенным ограничением. Это указывает на то, что лица, принимающие решения, регулирующие органы, исследователи и компании по-прежнему располагают ограниченными знаниями о том, как дезинформация, создаваемая GenAI, действует во многих других регионах. В докладе прямо заявляется, что это создает значительный пробел в знаниях за пределами тех немногих стран и языков, по которым публикуются материалы на эту тему. Этот пробел имеет важное значение для многоязычных обществ, трансграничного регулирования и усилий по разработке эффективных защитных механизмов, охватывающих различные языки, культуры и политические системы.

Кроме того, доказательная база ограничена быстрыми темпами технологического развития. Во многих экспериментах тестируются старые системы, такие как GPT-2, или ранние инструменты, подобные FaceSwap. В докладе содержится предупреждение о том, что научные исследования часто отстают от современных технологий. В результате политические решения иногда основываются на данных, полученных при изучении более слабых систем по сравнению с теми, которые в настоящее время используются в открытом доступе. Поэтому независимые исследования, изучающие более широкие контексты, являются практической необходимостью для обеспечения того, чтобы регулирование и управление соответствовали уровню развития технологий, формирующих общественный дискурс.

РЕЗУЛЬТАТ 3. ПРЕВЕНТИВНЫЕ КОРРЕКТИРОВКИ МОГУТ ПОМОЧЬ

Превентивная корректирующая информация включает краткие просветительские предупреждения, напоминания или побуждения, а также иные методы психологической «прививки» пользователей до контакта с потенциально вводящим в заблуждение контентом. Например, некоторые вмешательства предоставляют краткие объяснения того, как дипфейки

делают возможной реалистичную манипуляцию видео, или напоминают пользователям о том, что системы GenAI могут допускать ошибки. Эти подходы направлены на улучшение способности пользователей оценивать достоверность контента.

В исследованиях, проведенных после 2020 года, объединенная оценка показывает, что превентивная корректирующая информация снижает восприятие точности, достоверности или принятие на веру дезинформации, созданной GenAI, при этом размер эффекта варьируется от малого до умеренного. Эти выводы относительно согласуются друг с другом в рассмотренных исследованиях: вариативность их результатов невелика, а эффекты наблюдаются среди различных групп населения и при различных планах эксперимента в этой подборке данных.

ОСНОВНОЙ ВЫВОД

Превентивная корректирующая информация является наиболее стабильно эффективным вмешательством.

Эти результаты указывают на то, что такие вмешательства, вероятно, останутся эффективными в ближайшей перспективе, хотя большая часть имеющихся данных основана на исследованиях ранних поколений систем GenAI.

В Обобщающем обзоре также подчеркивается важная оговорка: положительное воздействие превентивной корректирующей информации проявляется тогда, когда людей просят оценить контент на предмет точности или достоверности. Данные менее однозначны в тех исследованиях, которые сосредоточены исключительно на распознавании дезинформации. В более ранних исследованиях дипфейков было обнаружено, что предупреждения иногда могли повышать уровень недоверия как к подлинным, так и к подвергшимся манипуляциям видеороликам. Это согласуется с более широкой проблемой в исследованиях и политике в области дезинформации: ложные нарративы и пропаганда часто направлены на то, чтобы заставить человека не доверять всей общедоступной информации. Следовательно, вмешательства должны быть разработаны таким образом, чтобы улучшить способность людей отличать достоверный контент от вводящего в заблуждение.

Превентивная корректирующая информация является одной из наиболее широко протестированных мер, которые могут быть относительно быстро внедрены в Интернете, особенно на платформах социальных сетей. Примеры включают примечания о том, что GenAI может допускать ошибки, объяснения того, как дипфейки делают возможными реалистичные манипуляции с видео, и краткие образовательные материалы о таких явлениях, как галлюцинации ИИ. В зависимости от научной дисциплины, такие меры называются рекомендациями, психологической «прививкой», превентивным разоблачением, праймингом или предупреждающей маркировкой.

РЕЗУЛЬТАТ 4. МАРКИРОВКА ЭФФЕКТИВНА ПРИ УСЛОВИИ ЕЕ ЕДИНООБРАЗНОГО ПРИМЕНЕНИЯ

ОСНОВНОЙ ВЫВОД

Маркировка контента остается многообещающей мерой, но только при условии ее ясности и единообразного применения.

Маркировка контента подразумевает использование коротких визуальных, текстовых или мультимодальных меток, прикрепляемых к информации. Эти метки указывают на то, что информация могла быть создана с помощью GenAI, подверглась изменениям или может вводить в заблуждение. Согласно результатам объединенного анализа, наличие меток связано с небольшим, но статистически значимым снижением воспринимаемой достоверности вводящей в заблуждение информации, созданной GenAI. Это обнадеживающий результат, однако он не является стабильным во всех случаях. В докладе отмечается значительно большая вариативность данных о маркировке по сравнению с данными о превентивной корректирующей информации.

Из-за ограниченного объема данных наблюдается значительная вариативность эффектов, а в некоторых исследованиях эффекты не обнаруживаются вовсе. Метки могут снижать воспринимаемую достоверность в одних ситуациях, но не оказывать такого воздействия в других. В данной оценке подчеркивается роль вероятных модерирующих факторов, включая дизайн метки, формулировку, способ представления, тип контента, условия эксперимента и источник метки. Иными словами, то, как метки разрабатываются и внедряются, имеет не меньшее значение, чем сам факт их использования.

Доказательная база в отношении маркировки также ограничена. Исследования в этой относительно небольшой выборке почти полностью сосредоточены на участниках из США. Это затрудняет понимание того, насколько эффективной окажется маркировка в различных языковых, культурных или правовых системах. Следовательно, правильный вывод заключается не в том, что маркировка неэффективна. Напротив, маркировка может быть полезной, но только в том случае, если компании уделяют внимание дизайну этих меток и тестируют их в различных контекстах, а не исходят из того, что они по умолчанию эффективны повсеместно.

ЗАКЛЮЧЕНИЕ

Данное Резюме для лиц, принимающих решения, освещает меняющуюся картину рисков. Дезинформация, созданная текстовым GenAI, представляет более серьезную угрозу убеждающего воздействия, чем дезинформация, созданная визуальным GenAI. Однако это не устраняет угрозы дипфейков и иных синтетических медиа. Это указывает на то, что лицам, принимающим решения, не следует позволять высокой заметности угрозы визуального обмана отвлекать внимание от нарастающей убеждающей силы ложных текстов, генерируемых ИИ.

В данном докладе подчеркиваются результаты предыдущих оценок Международной группы экспертов по информационной среде (IPIE) [3], [4] и предлагаются две практические меры реагирования. Превентивная корректирующая информация является наиболее устойчивым и надежно подкрепленным доказательствами вмешательством в имеющейся доказательной базе, особенно когда она предоставляется до контакта с дезинформацией и способствует формированию более взвешенных оценочных суждений. Маркировка контента также может быть полезной, однако ее эффективность варьируется и зависит от контекста. Метки наиболее эффективны, когда они тщательно разработаны и применяются ясно и единообразно.

Последний пункт касается самой доказательной базы. Большая часть мира остается за рамками текущих исследований. Это ограничивает возможности лиц, принимающих решения, платформ и разработчиков моделей ИИ общего назначения в понимании того, как эти риски распространяются в различных языковых, культурных и медийных средах. Поэтому независимые исследования, непрерывный синтез данных и расширение доступа исследователей к данным платформ и моделей имеют критически важное значение. Без этого государственная политика будет оставаться реактивной, неполной и слишком узконаправленной в условиях быстро развивающейся информационной среды.

СПИСОК ЛИТЕРАТУРЫ

[1] International Panel on the Information Environment [I. Trauthig, P. N. Howard, S. Valenzuela (eds.)], “The Role of Generative AI Use in 2024 Elections Worldwide,” [«Роль использования генеративного ИИ на выборах по всему миру в 2024 году»], Zurich, Switzerland: IPIE, 2025. Technical Paper, TP2025.2, doi: 10.61452/HZUE9853.

[2] International Panel on the Information Environment [A. Herasimenka, S. Valenzuela, S. Boulianne, F. Esser, L. M. Given, S. Lewandowsky, E. M. Navarro-López, P. N. Howard (eds.)], “Effects of Misinformation Produced with Generative AI and its Countermeasures: A Meta-Analysis of Experimental Scientific Evidence,” [«Последствия воздействия дезинформации, созданной генеративным ИИ, и меры противодействия ей: метаанализ экспериментальных научных данных»], Zurich, Switzerland: IPIE, 2026. Synthesis Report, SR2026.2, doi: 10.61452/UGTR3022.

[3] International Panel on the Information Environment, “Countermeasures for Mitigating Digital Misinformation: A Systematic Review,” [«Меры противодействия цифровой дезинформации: систематический обзор»], IPIE, Zurich, Switzerland, SR2023.1, July 2023. [Электронный ресурс]. Режим доступа: https://www.ipie.info/research/sr2023-1

[4] International Panel on the Information Environment, “Platform Responses to Misinformation: A Meta-Analysis of Data,” [«Ответные меры платформ на дезинформацию: метаанализ данных»], IPIE, Zurich, Switzerland, SR2023.2, July 2023. [Электронный ресурс]. Режим доступа: https://www.ipie.info/research/sr2023-2

БЛАГОДАРНОСТИ

Участники работы

Авторы-составители: Александр Герасименко (научный консультант, Великобритания), Себастьян Валенсуэла (директор по науке IPIE и председатель Комитета по науке и методологии, Чили), Шелли Булианн (член Комитета по науке и методологии IPIE, Канада), Франк Эссер (член Комитета по науке и методологии IPIE, Швейцария), Лиза М. Гивен (член Комитета по науке и методологии IPIE, Канада/Австралия), Стефан Левандовски (член Комитета по науке и методологии IPIE, Австралия/Великобритания), Ева М. Наварро-Лопес (член Комитета по науке и методологии IPIE, Испания/Великобритания/Мексика), Филип Ховард (президент и генеральный директор IPIE, Канада/Великобритания). Научные ассистенты: Анна Джордж и Сяньлинчэнь Ван. Независимое общее рецензирование: Джордж Георгаракис и Матиас Харрер. Дизайн: Доменико Ди Донна. Литературное редактирование: Беверли Сайкс. Мы выражаем искреннюю признательность за поддержку сотрудникам Секретариата IPIE в лице: Лолы Гимферрер, Джессики Голд, Виктории Шульц, Донны Сеймур, Анны Стендер и Алекса Янга.

Финансирующие организации

Международная группа экспертов по информационной среде (IPIE) с благодарностью отмечает поддержку своих финансирующих организаций. Полный список партнёров по финансированию размещён на сайте www.ipie.info. Любые мнения, выводы, заключения или рекомендации, содержащиеся в данном докладе, отражают позицию IPIE и не обязательно совпадают с точкой зрения финансирующих организаций.

Заявление о конфликте интересов

Доклады IPIE разрабатываются и рецензируются глобальной сетью ассоциированных исследователей и научных консультантов, которые образуют профильные научные группы и коллективы авторов. Все участники и рецензенты заполняют заявления о конфликте интересов, которые проверяются IPIE на соответствующих этапах работы.

Рекомендуемый формат цитирования

Резюме для лиц, принимающих решения, издаваемое Международной группой экспертов по информационной среде (IPIE), представляет собой обобщённую сводку о текущем состоянии знаний и адресовано широкой аудитории. В *Обобщающем докладе* IPIE используются методы научного метаанализа, систематические обзоры и другие инструменты для агрегирования данных, обобщения знаний и формирования научного консенсуса; данный документ предназначен для экспертной аудитории.

Технический документ IPIE посвящён конкретным вопросам методологии либо содержит анализ государственной политики применительно к отдельной проблеме нормативно-правового регулирования. Все доклады доступны на веб-сайте IPIE (www.IPIE.info).

Данный документ следует цитировать следующим образом:

International Panel on the Information Environment [A. Herasimenka, S. Valenzuela, S. Boulianne, F. Esser, L. M. Given, S. Lewandowsky, E. M. Navarro-López, P. N. Howard (eds.)], "Responding to Generative AI Misinformation: Results from a Scientific Meta-Analysis," [«Реагирование на дезинформацию, создаваемую генеративным ИИ: результаты научного метаанализа»], Zurich, Switzerland: IPIE, 2026. Summary for Policymakers [Резюме для лиц, принимающих решения], SFP2026.2, doi: 10.61452/NLZH2351.

Информация об авторских правах

ОБ IPIE

Международная группа экспертов по информационной среде (IPIE) — независимая глобальная научная организация, призванная обеспечивать наиболее практически применимые научные знания об угрозах мировой информационной среде. Миссия IPIE, базирующейся в Швейцарии, состоит в предоставлении лицам, принимающим решения, представителям отрасли и гражданскому обществу независимых научных оценок состояния глобальной информационной среды путём организации, оценки и продвижения исследований с общей целью улучшения глобальной информационной среды. Сотни исследователей со всего мира вносят свой вклад в создание докладов IPIE.

Для получения дополнительной информации обращайтесь в Международную группу экспертов по информационной среде (IPIE), secretariat@IPIE.info. Seefeldstrasse 123, P.O. Box, 8034 Zurich, Switzerland.

Международная
группа экспертов по
информационной
среде

Seefeldstrasse 123
P.O. Box 8034 Zurich
Switzerland