IPIE

International Panel on the Information Environment

Confronting Misinformation Produced with Generative AI

A Meta-Analysis of Experimental Scientific Evidence

Synthesis Report 2026.2

DOI Number: 10.61452/UGTR3022

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

IPIE
International Panel on the Information Environment

Confronting Misinformation Produced with Generative AI

A Meta-Analysis of Experimental Scientific Evidence

Synthesis Report 2026.2

How to cite:
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.)], "Confronting Misinformation Produced with Generative AI: A Meta-Analysis of Experimental Scientific Evidence," Zurich, Switzerland: IPIE, 2026. Synthesis Report, SR2026.2, doi: 10.61452/UGTR3022.

SYNOPSIS

This report synthesizes experimental evidence on the effects of generative artificial intelligence (GenAI) misinformation on individuals' ability to evaluate information and the effectiveness of interventions designed to mitigate its influence. It draws on a meta-analysis of experimental studies (60 effects from 24 publications and 33,801 participants) published between 2018 and 2025.

The findings indicate a divergence in public responses to different modalities of GenAI-produced misinformation (referred to in the report as GenAI misinformation), alongside substantial heterogeneity across studies. More recent research tends to report higher perceived accuracy and credibility of textual GenAI misinformation, whereas individuals' perceptions of visual GenAI misinformation—primarily deepfakes—have become more skeptical over time. These patterns are associated with the time of data collection rather than model capacity and are accompanied by wide variation in effect sizes across contexts.

The report also evaluates two countermeasures: preventive corrective information and content labeling. In studies conducted after 2020, corrective information yields small-to-moderate (0.17 to 0.43 standard deviation units) reductions in the perceived accuracy or credibility of GenAI misinformation. Pooled estimates in this subset are significant and comparatively stable, though prediction intervals show substantial variation, especially when outcomes focus on detection rather than evaluative judgment.

Content labeling interventions vary more than corrective information. Although labels are associated with modest average reductions in perceived credibility, effect sizes vary widely across modalities, designs, presentations, and contexts, and prediction intervals often include null effects.

Key conclusions from the synthesis are as follows:

1. Textual GenAI misinformation may currently pose greater persuasive risks than visual GenAI misinformation.
2. Preventive corrective information appears to be more a consistently effective countermeasure.
3. Content labeling policies yield highly variable outcomes and require careful design and evaluation.
4. The available evidence base is geographically narrow and methodologically constrained, with most studies focusing on populations in Western Europe and North America.

Based on these findings, policy responses should consider the following:

- Prioritize addressing text-based GenAI misinformation.
- Treat preventive corrections as a core intervention.
- Approach content labeling cautiously

CONTENTS

SECTION 1. INTRODUCTION

Generative artificial intelligence (GenAI) systems are capable of producing text, images, audio, and videos that humans at times have a hard time identifying as artificially generated. Examples of such systems are large language models (LLMs) like GPT and Gemini, which generate text; image synthesis models such as DALL-E and Midjourney; and video generation systems like Runway and Sora. However, GenAI-produced content may also be misleading and harmful, and it can be turned into "fake news," "rumors," "computational propaganda," or "deepfakes" [1], [2], [3]. In this Synthesis Report (2026.2), we refer to these phenomena by using the umbrella term "misinformation," which may or may not imply intention (see the full definition in Table 1). Moreover, the source of deceptive content often remains obscure when encountered on social media, the primary channel for digital misinformation [4], [5].

This report summarizes experimental evidence on the effects of exposure to GenAI-produced misinformation on individuals, as well as the countermeasures most effective at improving individuals' ability to evaluate it. Such misinformation may or may not be more persuasive than comparable misinformation that is not produced by GenAI [6], [7]. However, GenAI can now rapidly produce large volumes of misleading text, images, audio, and video with readily accessible tools [7], [8]. This content can be well-tailored to specific audiences and designed to mimic authentic communication, making it more challenging for users to evaluate its accuracy.

Given the potential impact of GenAI on democracy and public policy, understanding how individuals process and respond to misinformation generated by GenAI is a pressing challenge. With these urgent concerns in mind, we focus specifically on misinformation produced by GenAI rather than by any other system that might be classified as artificial intelligence in broad terms (see the full definition of GenAI in Table 1).

Table 1. Definitions of Key Concepts.

Concept	Definition	Source(s)
Countermeasure	Mitigating solutions to address GenAI-produced misinformation, such as content labeling and corrective information.	[1], [9][9]
Generative artificial intelligence	An algorithm, a system, or a piece of software designed to produce an output, especially text or images, "previously thought to require human intelligence, typically by using machine learning to extrapolate from large collections of data."	[26]
Information evaluation	Assessing information "credibility," "trustworthiness," "bias," "accuracy," and other similar properties of information.	See Appendix B for the full list of sources.
Misinformation	Misleading information and a range of related phenomena, including—but not limited to—disinformation, fake news, propaganda, rumors, credibility of information, manipulation, hallucination, and deepfakes.	[1]

To collect and compare the most reliable evidence, we focus on summarizing quantitative, experimental evidence originating in publications reporting results of randomized controlled trials (RCTs). These trials tested two types of effects: exposure to GenAI-produced misinformation and countermeasures aimed at reducing its influence. Experimental designs can be compared systematically, and the data they report certainly support inferential inquiry. Other methods, including qualitative approaches, can also yield valuable results, but synthesizing such evidence falls outside the scope of this report.

This Synthesis Report is based on 60 RCT effect estimates reported in 24 scientific publications. In comparison, a previous IPIE report on misinformation (SR2023.1) analyzed 43 effect estimates reported in 18 publications [9]. This report relies on effect estimates that are conceptually similar enough to warrant summarizing them in a meta-analysis.

Our current sample is derived from almost 7,000 peer-reviewed journal articles, book chapters, and conference proceedings spanning multiple disciplines, including media and communication, psychology, human-computer interaction,

political science, and related fields. We use a meta-analysis methodology to pool effect estimates reported in these publications and summarize the effects of GenAI-produced misinformation.

The evidence in this review is limited to quantitative experiments due to the nature of the meta-analysis method. Nevertheless, our framework is based on a large set of qualitative and quantitative literature that has been analyzed and summarized in previous publications. For example, we relied on the classification of countermeasures designed to address the effects of GenAI-produced misinformation on individuals (see definitions in Table 1) first refined in a previous IPIE report [1].

We employed a broad conceptual approach to capturing a variety of publications for our review, adopting flexible definitions. When selecting these publications, we carefully assessed each against the more specific methodological and empirical eligibility criteria outlined in Appendix C, Table C2. For example, some scholars consider "AI hallucination" part of the phenomenon of misinformation, while others are critical of this view [10]. However, a definition of misinformation that included "hallucination" as a keyword enabled us to discover at least one study that equated hallucination with misinformation and presented evidence relevant to our scope [11]. We reviewed this study's definitions and found they aligned closely with those of other studies included in our sample (see details on evidence selection in the Methods section). This example illustrates why a broader approach is beneficial for identifying relevant studies in such a large, cross-disciplinary review.

Research Questions

Although a growing body of studies has examined human responses to misinformation, much of this research has focused on manually crafted misleading content. Evidence on the psychological and behavioral effects of GenAI-produced misinformation remains limited, with individual studies differing in design, measures, and findings. A synthesis is needed to establish the magnitude and consistency of these effects across different contexts.

The first research question concerns what happens when a person is exposed to a misleading video or text created with GenAI. Such a video or text, for example, may include a GenAI-produced story claiming that a well-known politician said something totally opposed to what they were previously known to stand for, or may include a massive number of intentionally created, misleading statements about climate change generated using a ChatGPT model [12]. GenAI models are evolving extremely rapidly. Accordingly, we treat them as moving targets and analyze changes in their persuasiveness over time. Specifically, we ask the following questions:

RQ1: What effects does exposure to GenAI-produced misinformation have on individuals' evaluation of information, and how do these effects vary across content modality (textual versus visual), model type, and time?

Practical interventions are needed to counter GenAI misinformation. Debates about policy, platform regulations, and public discussions all hinge on whether countermeasures—such as labeling AI content, preemptive warnings, or accuracy nudges—can mitigate harms. Existing studies provide promising but sometimes contradictory evidence about the effectiveness of such interventions. Without systematic aggregation, it is unclear which countermeasures are robust and which are context-dependent. This gap motivates our second research question:

RQ2: Which countermeasures are most effective at improving individuals' ability to evaluate GenAI-produced misinformation?

By addressing these two questions through the meta-analysis of the latest peer-reviewed scientific research, our study contributes evidence that is both timely and policy-relevant. It moves beyond individual findings to identify consistent patterns, clarify effect size estimates, and highlight areas where further research is most needed.

In what follows, we present the conceptual framework, discuss the main results, and conclude with a summary of the findings.

SECTION 2. CONCEPTUALIZATION

This report extends previous IPIE reports that focused on strategies for mitigating misinformation acceptance [1], [9]. The earlier reports showed that content labeling and preventive corrective information may mitigate an individual's reliance on misinformation more effectively than other countermeasures (see SFP2023.1, SR2023.1, and SR2023.2). Less is known about the effects of these measures on GenAI-produced content. One aim of this report is to address this gap.

Before exploring the effects of GenAI-produced misinformation, such as deepfakes, it is important to determine whether this content actually persuades people or simply leaves them confused. Addressing this question is another aim of the current report. The experimental evidence is mixed. Some studies have shown that internet users are often unsure how to interpret deepfakes and similar GenAI-produced misinformation, responding with uncertainty rather than believing them [13]. In contrast, other experiments have reported that GenAI misinformation can be persuasive [7], but others find the opposite [14].

The scholarship highlights two realms of GenAI-produced misinformation:

1. Textual GenAI misinformation: Since 2021, when GPT-3 LLM became widely available, there has been a rapid improvement and growth in language- and text-focused systems known as LLMs. Their outputs are often hard to distinguish from human-produced content. One study found that "almost all LLMs tested released since 2022 produce election disinformation operation content indiscernible by human evaluators over 50% of the time" [15]. On average, the persuasiveness of such LLMs has increased over the years as models have become more complex and sophisticated, though their accuracy may be decreasing [16]. However, many other studies have shown that GenAI-produced misinformation may not be particularly persuasive, for example, [17], [18].

2. Visual GenAI misinformation: The problem of visual GenAI misinformation emerged well before the public release of GPT-3 [13]. This type of misinformation involves the production or modification of images, audio, or video. The modifications can contain varying degrees of misleading content, ranging from decontextualized images or cheapfakes to the most potent format—deepfakes. However, the effects of deepfakes and similar instances of GenAI-produced visual misinformation on individuals remain unclear [19].

While the broader literature includes audio-based misinformation, such as voice cloning, the available experimental evidence is limited and does not support a separate analysis. As a result, this report focuses on textual and visual modalities, with visual content primarily represented by images and video.

Another crucial variation in the literature concerns the type of model used to generate misinformation. Some studies have examined earlier models, such as GPT-2, and early deepfake software such as Faceswap. Others have employed more recent models, such as GPT-4, which are trained on larger datasets and have undergone numerous cycles of improvement, including imposing restrictions on algorithms to prevent the production of misinformation.

Finally, mitigating strategies to help individuals distinguish between misleading information and fact-based statements vary significantly. Multiple strategies are proposed, including reliance on external information authentication sources, multi-year information literacy programs, and even encrypted digital signatures attached to content.

A previous IPIE meta-analysis (SR2023.2) identified corrective information and content labeling as effective measures for improving an individual's ability to identify misinformation on social media platforms [9]. However, evidence for this conclusion comes from studies published before 2023, when GenAI-produced textual misinformation became more widespread. A recent systematic literature review concluded that "many proposed countermeasures remain either untested or lack real-world validation" in relation to the problem of AI and misinformation

[20]. This report analyzes the effectiveness of corrective information and content labeling in addressing GenAI-produced misinformation.

SECTION 3. METHODS

Identification of Publications

We searched for publications in two key academic databases, Scopus and Web of Science. These databases offer a valid instrument for evaluating journal articles, conference papers, books, and book chapters in the social sciences and humanities and in relation to human-computer interaction [21] and have been utilized in past scholarship that posed similar questions [22]. The publications include qualitative work in humanities journals and book chapters as well as synthetic-data-focused studies appearing in peer-reviewed computer science conference proceedings. These databases also offer tools for efficiently extracting the extensive bibliometric information required for a study of this scale.

We searched the databases using the search terms shown in Table C1 in Appendix C. We focused on publications about GenAI and misinformation, so we looked for synonyms of "generative AI" and "misinformation" in a publication's abstract, title, or keywords. Hence, in our initial review, we included publications that use any of the umbrella terms for misinformation (such as "misinformation," "misleading information," "disinformation," "propaganda," "fake news," or "rumors") in the main research questions. To address the ambiguity of terms related to "AI," we included the names of several leading generative AI models as search terms (see Table C1 in Appendix C), drawing from the Leaderboard ranking, an open-source, community-based list available at the time of data collection [23]. We then cross-checked this approach against the search terms used in recent systematic reviews focused on generative AI models [24] to ensure comprehensive coverage.

Of several search term combinations tested, we selected the set that yielded the most relevant studies. The final search request consisted of six terms for "misinformation" and 15 terms for "Generative Artificial Intelligence."

We included works published between 1 January 2017 and 10 April 2025. Additionally, in July 2025, we asked the IPIE affiliates (approximately 450

researchers) to suggest any other relevant publications for our review, resulting in 12 additional manuscripts. Some of these were not indexed in the two databases we were using.

Gray literature, including non-peer-reviewed conference abstracts, presentations, and preprints, was excluded. This approach strikes a balance between incorporating the most recent research and maintaining methodological consistency. Peer-reviewed research in media and communication, psychology, human-computer interaction, political science, and related fields forms the core of the evidence analyzed here.

Excluding gray literature may mean that some very recent findings—particularly early results disseminated via platforms such as arXiv—are not captured in this review. This approach to early results dissemination is more common in computer science than in the social scientific literature that is the focus of this review. While gray literature does not reflect lower evidentiary standards, it is difficult to identify and sample systematically because it is dispersed across heterogeneous and incompletely indexed sources. Including such materials as primary evidence would increase the risk of selection bias and reduce comparability across studies.

After finalizing the selection from each database (NScopus = 5,777, NWOS = 3,687, Nexpert = 12), we merged the results and removed duplicates, resulting in a total of 6,952 publications. This corpus, in addition to standard peer-reviewed journal articles, included 24 book chapters, 313 early-access publications, and 318 conference proceedings papers.

Eligibility Screening

To be eligible for inclusion, publications needed to satisfy five requirements derived from previous IPIE reports; they are defined below (for further details, see Appendix C Table C2).

1. Empirical—The publication includes an experiment or analysis of quantitative or qualitative data that generates evidence-based claims [25].

It must not be a purely opinion or position paper, although position papers with substantial empirical components are eligible;

2. Method—The publication uses an RCT or quasi-experimental design;
3. Misinformation— The publication discusses any aspects of misleading information and a range of related phenomena: misinformation, disinformation, fake news, propaganda, rumors, "credibility" of information, manipulation, and hallucination [1];
4. Generative AI—An algorithm, a system, or a piece of software designed to produce an output, especially text, "previously thought to require human intelligence, typically by using machine learning to extrapolate from large collections of data" [26];
5. Effect on individuals or Countermeasure—A publication must report the impact, effect, or consequences of exposure to GenAI-produced misinformation on human subjects or the implementation of a countermeasure designed to address the effects of GenAI-produced misinformation on individuals.

Three researchers (the Consulting Scientist and two Research Assistants; see the Contributors section) ran four rounds of pilot coding tests on approximately 150 records from our publication corpus to stress-test and refine the coding scheme used to synthesize the research evidence (see Table C2 in Appendix C). Throughout this process, disagreements about coding decisions were discussed and resolved collaboratively, resulting in a finalized coding scheme [27].

Following the pilot coding, the three researchers read the titles, abstracts, and keywords of 250 additional publications from the publication corpus to assess study eligibility based on the criteria listed above, in consultation with each other. In some cases, publications contained all relevant search terms but did not actually address misinformation, lacked any empirical component, were unrelated to GenAI, or did not report the effects of misinformation exposure or countermeasure implementation.

We classified all selected studies as RCTs. Such studies randomly assign individuals to receive a treatment or a control, while controlling for additional covariates to minimize confounding. No quasi-experimental studies met our inclusion criteria, and we excluded observational, correlational, qualitative, and descriptive studies.

We ran additional method-focused searches within our sample across abstracts, keywords, and titles, searching for expressions that might identify RCTs or quasi-experimental studies. We verified the results of these additional searches using a Gemma 2 detector model, an LLM by Google [28], [29], [30], with a prompt presented in Table C3 in Appendix C.

We used these searches to identify potentially eligible publications from yet-unevaluated publications. A human coder screened those publications that one of these approaches marked as potentially eligible for inclusion in our final sample. A manual review of 1,253 publications, or 18% of the sample, was conducted, and the rest were filtered out during additional keyword searches and LLM verification. Human coder evaluation of all the GenAI-excluded publications in line with the protocol detailed in Table C3 in Appendix C showed that the LLM verification incorrectly excluded 1.56% of the publications.

Inclusion: Synthesis and Analysis

After the titles, abstracts, and keywords were analyzed, 87 publications met the eligibility criteria. Working in pairs, the three coders then read the selected publications in full and extracted relevant data on the sample and reported effect sizes, following the Codebook (Table C2 in Appendix C).

The dependent and independent variables from each study are listed in individual summaries in Section 4 of this report and in Online Supplemental Appendix B. If a study provided statistics allowing for the calculation of pooled effect sizes, it was included in the meta-analysis. When standard errors were not available, they were calculated using an exact p value. We adjusted the effect sizes in the same direction to ensure a consistent conceptual meaning across all studies. Studies

that did not report effect sizes and standard errors in line with recommended practices were excluded from further analysis [31].

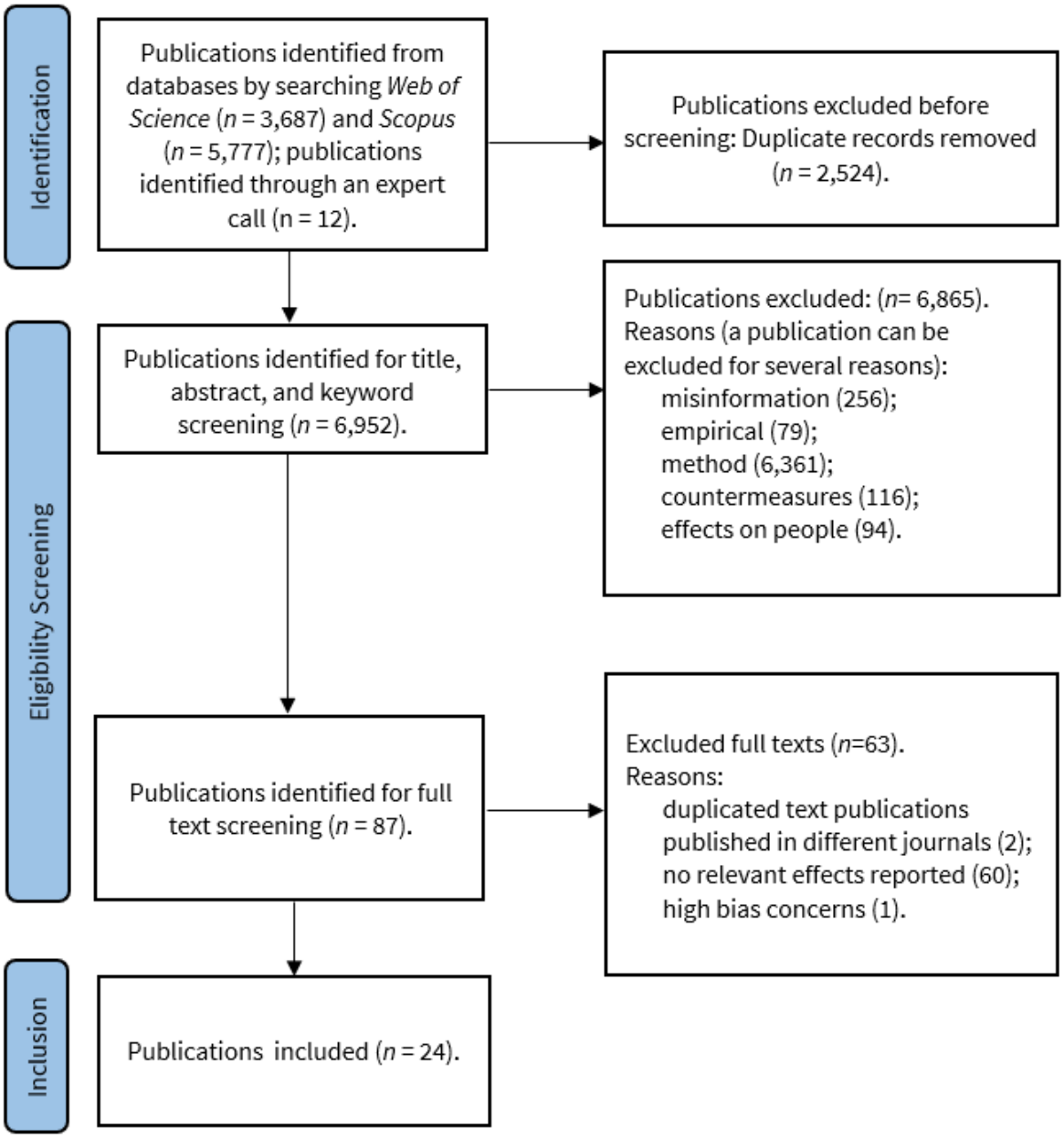
Using this meta-analysis approach, 60 effect estimates were extracted from 24 peer-reviewed publications (37 misinformation effects and 23 countermeasure effects). Appendix B provides the complete list of reviewed publications, and Appendix C Table C4 details which publications contributed to each section of this report. According to the Cochrane guidelines, this number of studies is sufficient for a meta-analysis. Figure 1 illustrates the sampling stages of the meta-analysis.

We calculated the pooled effect estimates using the random-effects size model, which accounts for variation across studies (heterogeneity) in the effect estimates [32]. The temporal models were fitted with a three-level random-effects meta-analysis using the metafor package in R [33]. The model specified sampling error nested within the study and included the year and month of data collection (experiment) as a moderator to capture temporal trends. Estimation was performed via maximum likelihood. Other data manipulations for summarizing effect estimates were performed using the bibliometrix package for R [34] and other R packages, as detailed in Online Supplemental Appendix C.

The studies included in the final sample aimed to identify the effects of exposure to GenAI-produced misinformation or the effects of a countermeasure intended to address misinformation, after controlling for a range of other factors and often comparing effects across two or more groups, with one typically a control group.

We assessed the risk of bias in the analyzed publications using a standard protocol, the Revised Tool for Risk of Bias in Randomized Trials (RoB 2) (see Online Supplemental Appendix A). As a result of this validity assessment, one publication was excluded due to multiple concerns linked to the risks of bias arising from the randomization process, the measurement of the outcome, and possible deviations from the intended interventions. Hence, the final sample consisted of 24 publications.

Figure 1. Sampling Stages of the Meta-Analysis.



Source: IPIE calculations based on data collected.

Note: Flow of publications is presented based on a standard design suggested by the PRISMA guidelines.

SECTION 4. FINDINGS

Effects of GenAI Misinformation

Comparing the Perception of Textual and Visual GenAI Outputs

Figures 2a and 2b summarize the estimated effects of exposure to GenAI-produced misinformation on information perception using a random-effects meta-regression of pooled Hedges' *g*—a standardized measure of effect size capturing the average difference between exposed and unexposed groups, expressed in standard deviation units and combined across multiple studies—plotted by output modality and data-collection date.

The results indicate diverging temporal patterns in responses to textual versus visual GenAI-produced misinformation. As shown in Figure 2a, more recent studies generally report increasingly positive effect sizes for textual outputs, suggesting that exposure to textual GenAI misinformation is associated with a higher level of perceived accuracy or credibility relative to reliable information. In contrast, Figure 2b shows a downward trend over time for visual GenAI misinformation, with studies collecting data after roughly 2021 more frequently reporting negative effect sizes, indicating a reduced level of perceived credibility. The overall summary of the effects we identified in the literature, including some not analyzed with meta-regression, is presented in Online Supplemental Appendix B, Figure B1.

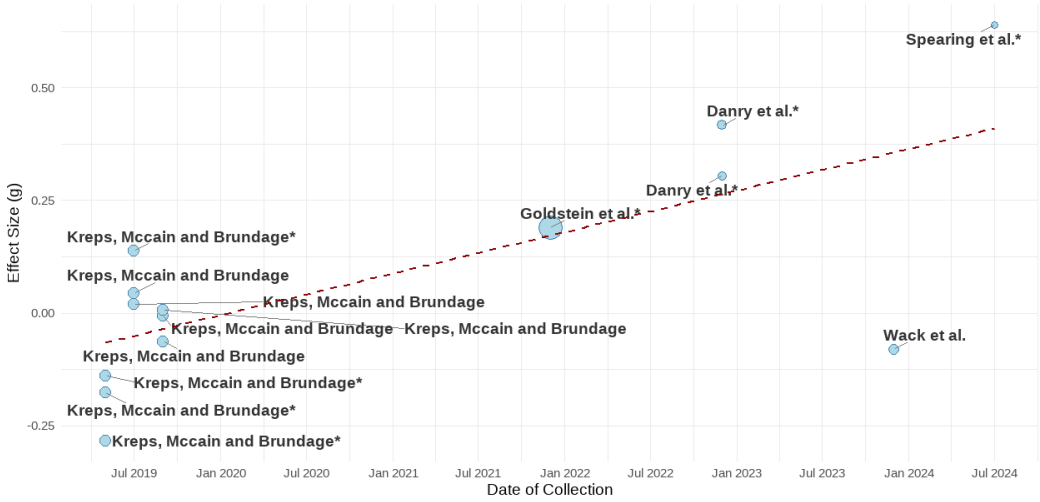
These shifts between 2020 and 2021 mark a broader transition in the availability and public prominence of GenAI technologies. It is important to note that the estimates reflect associations with the timing of the data collection rather than Gen AI models' capabilities. Across both modalities, between-study heterogeneity is substantial, with effect sizes varying widely in magnitude and direction. The meta-regression trends should therefore be interpreted as descriptive summaries of highly heterogeneous evidence rather than as indicators of uniform effects, a point discussed further in this section.

Textual Outputs

On average, textual GenAI misinformation outputs show an upward temporal trend in terms of perceived realism, accuracy, and related phenomena in the pooled meta-analytic model. The meta-regression shows that effect sizes increase as a function of the date of data collection, with a statistically significant positive slope for experiments conducted between 2019 and 2024 ($\beta = 0.08$, indicating a modest increase in effect size over time; 95% CI: [0.02; 0.15], indicating the range of values consistent with the data; $p = 0.01$, indicating that this pattern is unlikely to be due to chance; see Figure 2a for the visualization of all non-pooled Hedges' *g* estimates).

The estimated variance components were τ^2 (Level 3) = 0.01 and τ^2 (Level 2) = 0.01, each indicating the extent to which results vary across and within studies, respectively. Accordingly, I^2(Level 3) = 50.10% of the total variance is attributable to differences between publications, while I^2(Level 2) = 40.01% reflects differences between experiments within the same publication. These values indicate

Figure 2a. Exposure Effects on Perceived Textual GenAI Misinformation by Data Type and Collection Period.



Source: IPIE calculations based on data collected.
Note: Effect sizes (*g*) indicate individuals' evaluations of GenAI-produced misinformation as "accurate," "credible," or similar relative to reliable information. Each datapoint represents an individual effect estimate reported in a publication; larger points indicate greater study weight. Estimates are derived from a random-effects meta-regression based on 14 effects from 5 publications. See Appendix B for full references. When not reported, dates of data collection were estimated (see Methods section).
*Statistically significant effect estimate.

substantial heterogeneity, meaning that the magnitude and direction of effects vary considerably across research contexts. At the same time, a large proportion of the variation arises from differences between experiments reported within the same publication.

One illustrative example comes from a publication showing that a larger version of GPT-2 (LLMs with 774M parameters), now considered outdated, can generate misinformation that individuals evaluate as more credible than verified content, whereas a smaller model (355M parameters) did not yield comparable effects [17]. While this finding does not establish a causal link between model size and perceived credibility, it is consistent with the idea that more complex language models can generate content that is more coherent, nuanced, and, in this study's context, believable.

For more capable language models released after 2020, such as GPT-4 compared to GPT-2, pooled estimates suggest the potential for stronger misinformation-related effects (see Online Supplemental Appendix B Figure B2). The pooled effect size for these later models is significant ($g = 0.28$). However, the associated prediction interval—a measure of uncertainty about the mean that also accounts for variation (heterogeneity) between studies, unlike the confidence interval—includes 0, indicating that such effects may not be observed consistently across contexts. Indeed, a recent study using GPT-4 found that individuals were still better at recognizing GenAI-produced misinformation than human-generated misinformation [6].

The most pronounced deviation from the overall temporal trend in the textual dataset is observed in the research conducted by Wack et al. [18]. This outlier likely reflects differences in experimental design and stimulus construction. Unlike most other text-focused experiments, which relied on laboratory-made LLM outputs, this study used real-world misinformation that journalistic investigations suggested may have been produced with LLMs. However, the origins of these materials could not be confirmed by their producers.

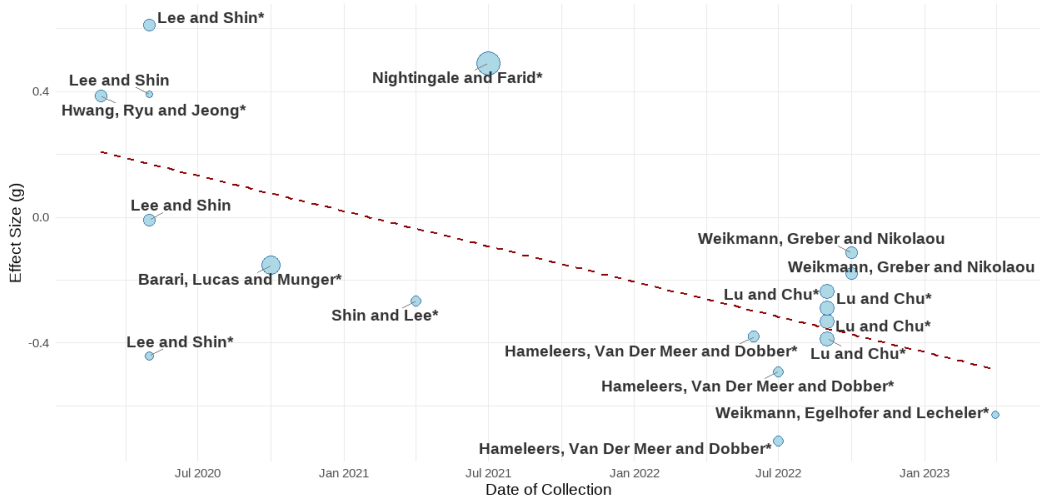
Visual GenAI Outputs

In contrast to the upward trend observed for textual outputs, visual GenAI-produced misinformation shows a statistically significant decline in the level of perceived credibility over approximately the same time period in the pooled meta-analytic analysis (β = -0.22, 95%CI: [-0.35; -0.10], p = 0.0003; see Figure 2b for non-pooled Hedges' *g* estimates). This negative slope indicates that, on average, more recent studies report lower levels of perceived realism or credibility for visual GenAI misinformation, such as deepfakes, relative to reliable information.

The estimated variance components were τ^2(Level 3) = 0.06 and τ^2(Level 2) = 0.00. In this model, most heterogeneity (I^2(Level 2) = 89.73%) is attributable to within-publication variation, while none of the heterogeneity is due to variation across experiments reported in different publications. Still, this result should not be interpreted as definitive evidence of the absence of between-study heterogeneity.

Figure 3 presents a meta-analysis of studies examining visual GenAI-produced misinformation using data collected after 2021. The pooled random-effects

Figure 2b. Exposure Effects on Perceived Visual GenAI Misinformation by Data Type and Collection Period.

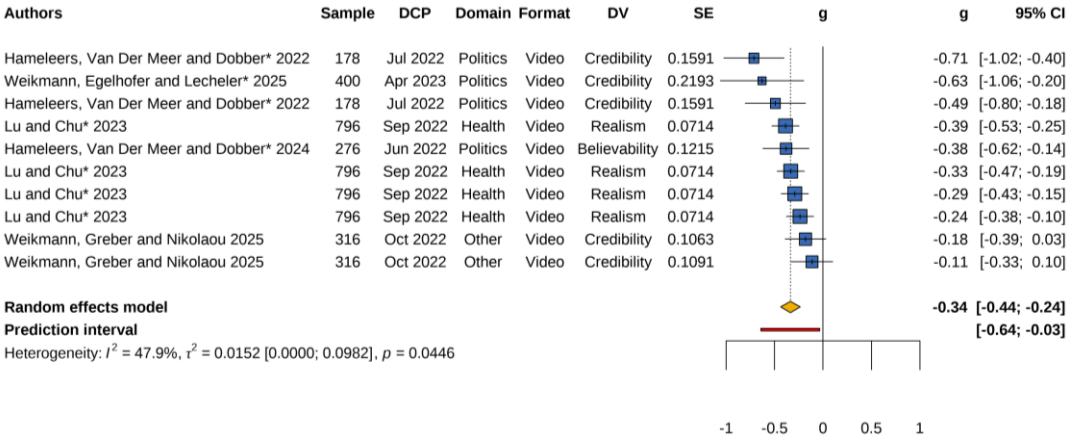


Source: IPIE calculations based on data collected.
Note: Effect sizes (*g*) indicate individuals' evaluations of GenAI-produced misinformation as "accurate," "credible," or similar relative to reliable information. Each datapoint represents an individual effect estimate reported in a publication; larger points indicate great study weight. Estimates are derived from a random-effects meta-regression based on 18 effects from 10 publications. See Appendix B for full references. When not reported, dates of data collection were estimated (see Methods section).
*Statistically significant effect estimate.

estimate shows a moderate and statistically significant decrease in the level of perceived "accuracy," "credibility," and similar evaluations (g = -0.34, 95% CI: [-0.44; -0.24]). This trend was observed across multiple domains, including politics and culture, and in studies with varying sample sizes and publication dates. The associated prediction interval (PI: [-0.64; -0.03]) suggests that, based on current evidence, future studies are expected to observe effects that are generally negative though potentially small in magnitude [35].

For the studies included in Figure 3, τ^2 was estimated at 0.01 (95% CI: [0.00; 0.09]), while I^2 was estimated at 47.9% (p = 0.04). Taken together, these estimates suggest a moderate dispersion of effects, with uncertainty around the true between-study variance. The evidence from publication bias diagnostics was mixed: some indicators suggest the presence of small-study bias, whereby smaller studies with larger effects are more likely to be published, potentially inflating pooled estimates, while other tests did not detect strong publication bias (see Methods and Online Supplemental Appendix A).

Figure 3. Exposure Effects on Perceived Visual GenAI Misinformation in Post-2021 Experiments.



Source: IPIE calculations based on data collected.

Note: Effect sizes (g) indicate individuals' evaluations of GenAI-produced misinformation as "accurate," "credible," or similar relative to reliable information. Data Collection Period (DCP) and Dependent Variable (DV) are provided. Effect sizes were calculated from random-effects meta-analysis using Hedges' g. τ and corresponding p were calculated using the Paule-Mandel estimator. Summary of 10 effects extracted from 5 publications.

*Statistically significant effect estimate.

Several underlying studies attribute difficulties in distinguishing authentic videos from deepfakes to factors such as trust in information sources [36], age (older participants were found to be more vulnerable to deepfakes than younger people) [36], [37], and reliance on visual cues, such as facial features, video background, or video quality, rather than contextual or social factors [37]. These explanations originate from individual studies and help contextualize the observed meta-analytic patterns but are not directly estimated in the present models. A key limitation of this subsample is its geographic concentration: with one exception, all studies focus on participants from the USA and the European Union. In the Limitations section below, we elaborate on the implications for future research of the restricted geographic coverage of the literature.

Countermeasures to address GenAI Misinformation

Preventive Corrective Information

Publications analyzing corrective information typically assess participants' evaluations of misinformation before and after exposure to preemptive educational forewarnings about potential deception. These interventions, sometimes described as "advice" or "priming," aim to improve individuals' ability to evaluate information by explaining common deceptive techniques used in GenAI-produced content and providing cues to identify misinformation [38].

Such forewarnings include reminders that GenAI can produce errors [8], explanations of how deepfakes enable realistic video manipulation [39], or brief educational materials on GenAI hallucinations [11]. These techniques form a subdomain of measures that are often also described as inoculation and media literacy [40], as they aim to influence individuals' evaluation skills before exposure rather than reactively correcting false claims. For a synthesis of the evidence on these interventions, see IPIE reports SFP2023.1, SR2023.1, and SR2023.2.

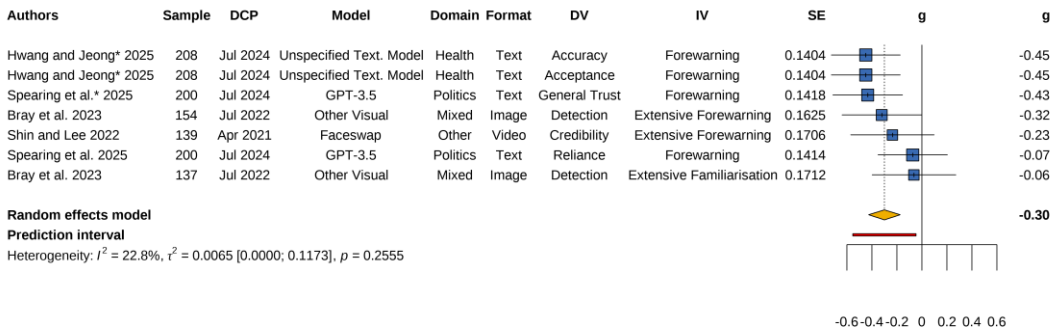
Figure 4 presents the results of the meta-analysis of studies examining preventive corrective information. In line with the emphasis on data-collection timing in

earlier analyses, this model includes only publications that collected data after 2020.

The pooled random-effects estimate indicates that exposure to preventive corrective information is associated with lower levels of evaluations of GenAI-generated misinformation as accurate, credible, or reliable (g = -0.30, 95% CI: [-0.43; -0.17], p < 0.0001). The corresponding prediction interval does not include zero, suggesting that negative effects are expected across a range of comparable future studies. Egger's test for funnel plot asymmetry was not statistically significant, and additional diagnostic tests did not suggest substantial publication bias or small-study effects within this subset (see Online Supplemental Appendix A).

Between-study heterogeneity in this post-2020 sample is low and non-significant, (I^2 = 22.8%, p = 0.25), indicating that variability across studies is consistent with chance. This suggests that the association between preventive corrective information and reduced perceived credibility of GenAI misinformation is consistent across the included study designs and contexts.

Figure 4. The Effect of Corrective Information on Misinformation Perception in Post-2020 Studies.



Source: IPIE calculations based on data collected.

Note: Effect sizes (g) indicate individuals' evaluations of GenAI-produced misinformation as "accurate," "credible," or similar relative to reliable information. Data Collection Period (DCP) and the main Independent Variable (IV) are provided. Effect sizes were calculated from random-effects meta-analysis using Hedges' g. τ and corresponding p were calculated using the Paule-Mandel estimator. Based on 7 effects extracted from 4 publications.
*Statistically significant effect estimate.

The results for the full sample, irrespective of the data collection period, are presented in Online Supplemental Appendix B Figure B4. While the pooled estimate remains negative (g = -0.22), uncertainty increases: the prediction interval includes 0, and heterogeneity is higher. These results highlight that the effectiveness of preventive corrective information is less consistent in earlier studies, especially those conducted before 2021.

The overall sample spans a range of populations and includes studies conducted in European countries (Austria, Germany, the Netherlands, and the UK), South Korea, and the USA, and individuals were recruited through platforms such as MTurk. Political misinformation constitutes the most common substantive focus. Within the subset of politics-focused studies, the pooled effect estimate is negative but inconclusive (g = -0.15, 95% CI: [-0.28; -0.02]), with substantial heterogeneity and a prediction interval that includes 0 (see Online Supplemental Appendix B Figure B7).

Experiments conducted between 2018 and 2020 generally report effects in the same direction as later research but with greater uncertainty and a higher level of heterogeneity (I^2 = 0.81; prediction interval includes 0; see Online Supplemental Appendix B Figure B6). An early study on deepfakes, for instance, found that exposure to corrective information led participants to distrust both authentic and manipulated videos [39].

The uncertainty of findings in earlier research appears to be partly related to differences in outcome operationalization. Several pre-2021 studies focus on individuals' ability to "detect" misinformation rather than on perceived credibility or accuracy. Within this subset, results are mixed: some studies report significant improvements in detection, while others find null effects, contributing to increased heterogeneity. Other works have argued that improving individual detection skills may be less effective than fostering trust in external sources of content authentication [37]. This highlights a conceptual distinction between detection and evaluative judgment. Consistent with this distinction, studies that did not operationalize outcomes in terms of detection exhibited a lower and

stronger, though still inconclusive, pooled effect estimate (g = -0.25; 95% CI: [-0.38; -0.12]; PI: [-0.61; 0.12]; see Online Supplemental Appendix B Figure B5.

Content Labeling

Content labeling refers to the use of brief visual, textual, or multimodal tags attached to information to signal that this information may be misleading or was generated using artificial intelligence [1]. This labeling can be automatically generated, moderator-generated, or user-generated. Examples of popular labels that researchers successfully tested include such tags as “AI-Generated” (“media that was generated using artificial intelligence”) [41], “Artificial” (“artificial content that has been edited or digitally altered”) [41], and “False information. This post has been reviewed by independent fact-checkers” [42]. In contrast to preventive corrective information, labeling interventions function primarily as direct warnings rather than as educational or skill-building measures. In the majority of the studies analyzed, warnings were typically displayed right before or during exposure to content, although some studies tested retroactive labeling applied after initial exposure [8].

The publications included in this analysis primarily examine how labels influence perceived believability and credibility of GenAI-produced misinformation. For instance, one study analyzing marks applied to images found that an “AI-generated” label was less effective than “manipulated” or “altered” labels [41].

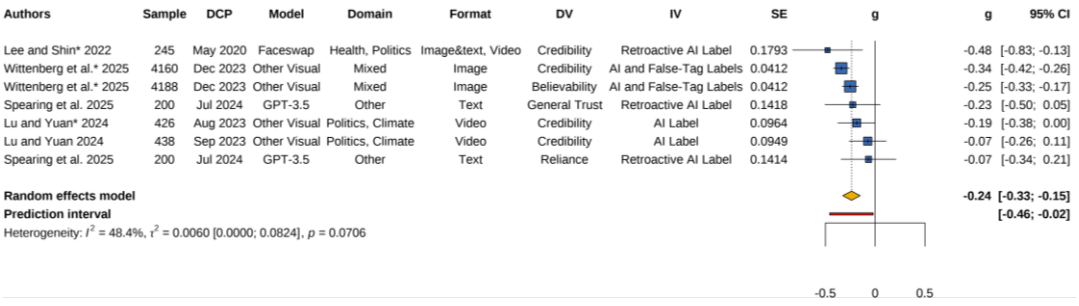
Figure 5 presents the results of the random-effects meta-analysis of content labeling interventions. The pooled effect indicates that the inclusion of a content label is associated with a small but statistically significant reduction in the perceived credibility of GenAI-produced misinformation (g = -0.24, 95% CI: [-0.33; -0.15], $p < 0.001$); after adjusting for publication bias, (g = -0.25 (95% CI: [-0.35; -0.16], $p < 0.001$). The analyzed studies include multimodal misinformation produced with a variety of models, including GPT variants and Faceswap. While image-focused studies tend to report larger negative effects than text-focused labeling, this pattern is descriptive and is not based on a formal analysis.

Between-study heterogeneity in this sample is substantial (I^2 = 48.4%), indicating variation in effect sizes across the studies. Consistent with this dispersion, the prediction interval is close to 0, suggesting that labeling interventions may not reliably reduce perceived credibility in some contexts. A sensitivity analysis showed that heterogeneity further decreased when one study involving a parody deepfake featuring former U.S. president Joe Biden was excluded (publication 16 in Appendix B). The authors of the paper acknowledged that factors such as the style of the deepfake and the personalities involved may have influenced the outcome. Additional variation in sample sizes was also observed across the analyzed studies.

The high level of heterogeneity observed in the content labeling literature suggests that contextual factors, such as label design, presentation, content type, and experimental setting, are likely to moderate effectiveness. In addition, the studies in this sample focus almost exclusively on participants in the USA, limiting the generalizability of the findings to other cultural or media environments.

Publication bias diagnostics did not reveal strong evidence of small-study effects in this subset (see Online Supplemental Appendix A and Figure A3). However, given the limited number of studies and the observed heterogeneity, the pooled estimate should be interpreted as a context-dependent average rather than as a uniform effect of content labeling.

Figure 5. The Effect of Content Labeling on Misinformation Perception.



Source: IPIE calculations based on data collected.

Note: Effect sizes (*g*) indicate individuals' evaluations of GenAI-produced misinformation as "accurate," "credible," or similar after exposure to a labeling countermeasure. Data Collection Period (DCP), Dependent Variable (DV), and the main Independent Variable (IV, a type of label) are provided. Effect sizes were calculated from random-effects meta-analysis using Hedges' *g*. τ and corresponding *p* were calculated using the Paule-Mandel estimator. Based on 7 effects extracted from 4 publications. *Statistically significant effect estimate.

Taken together, these findings suggest that while both preventive corrections and labeling are associated with reductions in the perceived credibility of GenAI misinformation on average, their effects are not uniform, and estimates should be interpreted as context-dependent rather than universally generalizable.

Limitations

Several methodological and evidentiary limitations should be considered when interpreting the results of this analysis. Many of these constraints are inherent in the current state of the literature and in the specific meta-analytic approach and methodology used in this literature.

Some pooled effect sizes included only a few estimates—no more than 7 in one case. There is no methodological consensus on the minimum number of estimates that are necessary: while some methodologists recommend including more, others do not impose a strict minimum and instead highlight the trade-off between the number of estimates and the power to detect the effects of a given size [43].

A similar observation is relevant for meta-analytic techniques. Our meta-analytic models are based on 14–18 estimates reported in 5–10 primary studies. Some guidelines advocate “at least 30 estimates of the effect size reported in at least 10 primary studies” [44], while others argue that it is best to prioritize a small number of highly theoretically relevant moderators over a larger, less relevant set [35]. Our emphasis on theoretical relevance partly explains the lower number of estimates included. Nevertheless, meta-regression based on fewer estimates will likely have lower statistical power, which means the results should be interpreted with caution.

Consequently, both meta-analytic models examining the effects of GenAI misinformation have low statistical power due to the small number of available studies. Regarding text-based experiments, substantial between-study heterogeneity persisted even with fixed predictors, and the confidence interval was close to zero, suggesting that observed relationships may not be relevant for

certain contexts. Our findings should be revisited when more studies become available.

It is also possible that relevant peer-reviewed literature exists outside the databases included in this review. In particular, gray literature, such as preprints, may offer valuable additional data but was not included in the review. However, most publications included were published in journals that belong to disciplines in which the practice of preprints is less common than, for example, in computer science. In other fields, such as human-computer interaction, preprints are more usual.

Despite efforts to broaden the scope of the literature, for example, by using databases to find publications that also index non-English journals, most of the analyzed publications focus heavily on Western European and North American contexts, which is a common issue in studies of GenAI misinformation [20]. Five non-English abstracts were evaluated, but none met the eligibility criteria. Consequently, the review was restricted to studying English-language full-text articles. This demonstrates a substantial gap in our knowledge about the phenomenon of GenAI misinformation beyond a few countries that invest heavily in relevant research.

This RCT-focused review captures a specific area of research concentrated in several disciplines. Findings from RCTs may not capture the full complexity of individuals' perceptions, interactions, and everyday experiences with GenAI misinformation. Integrating qualitative and observational studies will be necessary to add these important nuances, and this report should be seen as a first step toward this integration.

As often happens with such broad and ambitious literature scoping, our results show high levels of among-study heterogeneity for the samples focusing on content labeling (48.8%). This may indicate that the specifics of how a label is applied matter more than the mere presence of a label. At the same time, a high level of heterogeneity is common in misinformation research and broader social

science due to the wide variation in methodologies and the characteristics of independent variables [9], [45].

Given the relatively small number of studies examined and the sometimes less-than-optimal quality of the evidence, it was difficult to pinpoint the exact causal sources of heterogeneity. In such cases, we also observed indications of small-study bias, which may partially account for this variation. For a more detailed discussion of this problem, see Appendix A and our earlier report [9].

SECTION 5. CONCLUSION

We reviewed how individuals, in our case predominantly those in Western Europe, the USA, and South Korea, perceive GenAI-generated misinformation and what mitigating measures can be put in place to help them evaluate such misinformation. The results show that, broadly speaking, textual GenAI misinformation is increasingly perceived by individuals as more credible than similar verified information, while visual deepfakes are perceived as less credible.

The growing perceived credibility and believability of textual GenAI misinformation is even more concerning, considering recent findings suggesting that chats or conversations with GenAI are potentially 50% more persuasive than reading static GenAI-generated messages [16]. Hence, our estimate of the potentially greater credibility of GenAI misinformation texts ($\beta = 0.08$) is very conservative, as most of the experiments we reviewed examined the effects of static rather than conversational messages.

At the same time, visual deepfake-style misinformation may also improve in the future, and the public may once again find such content more credible. The trajectory of visual misinformation may therefore converge with the upward trajectory observed in textual models.

Preventive corrective information and content labeling were confirmed as two potentially effective measures for improving individuals' ability to evaluate GenAI-generated misinformation. Multiple experiments in our sample demonstrated the effectiveness of these two measures, though more evidence is needed to determine whether they remain effective across different contexts. For example, recent evidence, which was not available until after we completed this review, also confirms that preventive corrective information is likely to remain effective across various contexts [46]. At the same time, the effectiveness of content labeling measures is context-dependent at the moment, and more research is needed to demonstrate their broader potential.

The literature was also not conclusive regarding the effect sizes of the proposed countermeasures. Deciding whether an effect is "large" is not straightforward in any discipline, and this becomes even more difficult when researchers combine studies from different disciplines. The effect sizes for two key countermeasures that were analyzed—0.3 for preventive corrective information and 0.24 for labeling—fall between "small (~0.2 s.d.)" and "medium (~0.5 s.d.)" according to Cohen's standards. Although this interpretation of the findings is commonly applied across fields, it was not intended to be absolute [47]. This means that the size of effects must be judged within the context of the field and the methods used in the relevant study. Even a 0.3 reduction in the credibility of misinformation across a platform of 1 billion users has a massive practical impact.

The low number of estimates used in some of the literature samples means that the results should be interpreted with caution. Low estimates reduced the statistical power of meta-analytic models, and our conclusions about the effectiveness of labeling remain uncertain due to high heterogeneity. Further work is required to track the field's development, update the present review, and build a larger sample of highly homogeneous experiments that ask similar questions about the effects of GenAI-produced misinformation and the effectiveness of countermeasures to address this misinformation.

Policy Implications

Policy responses should prioritize text-based GenAI misinformation

The evidence presented in this report indicates that text-based GenAI misinformation is already perceived as highly credible. In some cases, it is even perceived as more credible than verified information. This does not dismiss the risks of visual GenAI-produced misinformation. However, the evidence suggests that interactive exchanges are substantially more persuasive than static messages, and conversational GenAI systems are increasingly becoming a key element of the global information environment [16]. This could make conversational, text-focused GenAI systems the most immediate and scalable

source of harm. Safeguards should extend explicitly to conversational GenAI systems.

Preventive corrective information should be treated as a core intervention

Across multiple experimental contexts, advance and explanatory warnings improve individuals' ability to evaluate GenAI-generated misinformation. Although effect sizes are modest by conventional standards, their potential impact at scale is substantial. Policymakers should therefore require or at least strongly encourage the integration of preventive corrective information into high-reach platforms and GenAI systems, particularly in domains with an elevated societal risk, such as climate change, public health, and elections.

Content labeling interventions are promising but require continuous evaluation

While labeling can improve users' ability to identify GenAI-generated misinformation, its effectiveness varies across contexts and label types. However, evidence in this area is less abundant than that regarding corrective information interventions. Before more research evidence emerges, regulatory frameworks should encourage iterative testing and refinement of this strategy.

Investment in independent evaluation and evidence synthesis is necessary

The limited number of estimates of the effects of GenAI misinformation and the high level of heterogeneity observed in parts of the literature indicate that current evidence remains incomplete, particularly regarding the effectiveness of countermeasures to mitigate GenAI misinformation. Policymakers should support coordinated, cross-disciplinary research efforts, replication studies, and regular updates to evidence reviews to ensure that governance approaches remain aligned with an evolving empirical base.

REFERENCES

[1] International Panel on the Information Environment, “Countermeasures for Mitigating Digital Misinformation: A Systematic Review,” IPIE, Zurich, Switzerland, SR2023.1, May 2023. [Online]. Available: https://www.ipie.info/research/sr2023-1

[2] A. Herasimenka, J. Bright, A. Knuutila, and P. N. Howard, “Misinformation and Professional News on Largely Unmoderated Platforms: The Case of Telegram,” Journal of Information Technology & Politics, vol. 20, no. 2, pp. 1–15, 2023, doi: 10.1080/19331681.2022.2076272.

[3] P. N. Howard, Lie Machines: How to Save Democracy From Troll Armies, Deceitful Robots, Junk News Operations, and Political Operatives. New Haven: London: Yale University Press, 2020.

[4] D. Caled and M. J. Silva, “Digital media and misinformation: An outlook on multidisciplinary strategies against manipulation,” J Comput Soc Sc, May 2021, doi: 10.1007/s42001-021-00118-8.

[5] Y. M. Rocha, G. A. de Moura, G. A. Desidério, C. H. de Oliveira, F. D. Lourenço, and L. D. de Figueiredo Nicolete, “The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review,” J Public Health (Berl.), Oct. 2021, doi: 10.1007/s10389-021-01658-z.

[6] A. Bashardoust, S. Feuerriegel, and Y. R. Shrestha, “Comparing the Willingness to Share for Human-generated vs. AI-generated Fake News,” Proc. ACM Hum.-Comput. Interact., vol. 8, no. CSCW2, p. 489:1-489:21, Nov. 2024, doi: 10.1145/3687028.

[7] J. A. Goldstein, J. Chao, S. Grossman, A. Stamos, and M. Tomz, “How persuasive is AI-generated propaganda?,” PNAS Nexus, vol. 3, no. 2, p. pgae034, Feb. 2024, doi: 10.1093/pnasnexus/pgae034.

[8] E. R. Spearing et al., “Countering AI-generated misinformation with pre-emptive source discreditation and debunking,” Royal Society Open Science, vol. 12, no. 6, p. 242148, Jun. 2025, doi: 10.1098/rsos.242148.

[9] International Panel on the Information Environment, “Platform Responses to Misinformation A Meta-Analysis of Data,” IPIE, Zurich, Switzerland, SR2023.2, May 2023. [Online]. Available: https://www.ipie.info/research/sr2023-2

[10] A. Shao, “New sources of inaccuracy? A conceptual framework for studying AI hallucinations,” Harvard Kennedy School Misinformation Review, Aug. 2025, doi: 10.37016/mr-2020-182.

[11] Y. Hwang and S.-H. Jeong, “Generative Artificial Intelligence and Misinformation Acceptance: An Experimental Test of the Effect of Forewarning About Artificial Intelligence Hallucination,” Cyberpsychology, Behavior, and Social Networking, p. cyber.2024.0407, Feb. 2025, doi: 10.1089/cyber.2024.0407.

[12] I. Trauthig, P. N. Howard, and S. Valenzuela, “The Role of Generative AI Use in 2024 Elections Worldwide,” International Panel on the Information Environment, Zurich, Switzerland, Technical Paper TP2025.2, 2025. Accessed: Dec. 24, 2025. [Online]. Available: https://www.ipie.info/research/tp2025-2

[13] C. Vaccari and A. Chadwick, “Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News,” Social Media + Society, vol. 6, no. 1, p. 2056305120903408, Jan. 2020, doi: 10.1177/2056305120903408.

[14] S. Barari, C. Lucas, and K. Munger, “Political Deepfakes Are as Credible as Other Fake Media and (Sometimes) Real Media,” The Journal of Politics, vol. 87, no. 2, pp. 510–526, Apr. 2025, doi: 10.1086/732990.

[15] A. R. Williams et al., “Large language models can consistently generate high-quality content for election disinformation operations,” PLOS ONE, vol. 20, no. 3, p. e0317421, Mar. 2025, doi: 10.1371/journal.pone.0317421.

[16] K. Hackenburg et al., “The levers of political persuasion with conversational artificial intelligence,” Science, vol. 390, no. 6777, p. eaea3884, Dec. 2025, doi: 10.1126/science.aea3884.
[17] S. Kreps, R. M. McCain, and M. Brundage, “All the News That’s Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation,” Journal of Experimental Political Science, vol. 9, no. 1, 2022, doi: 10.1017/XPS.2020.37.
[18] M. Wack, C. Ehrett, D. Linvill, and P. Warren, “Generative propaganda: Evidence of AI’s impact from a state-backed disinformation campaign,” PNAS Nexus, vol. 4, no. 4, p. pgaf083, Apr. 2025, doi: 10.1093/pnasnexus/pgaf083.
[19] S. Kang and K. Valadez, “Deepfakes’ Cognitive, Emotional, and Behavioral Impact: A Systematic Review and Meta-Analysis of Individual Responses,” Journalism & Mass Communication Quarterly, vol. 102, no. 4, pp. 958–992, Dec. 2025, doi: 10.1177/10776990251357294.
[20] M. F. Grub and E. Humprecht, “Generative AI and Disinformation| Defining the Role(s) of AI in Disinformation Research—A Systematic Review,” International Journal of Communication, vol. 19, Nov. 2025.
[21] M. Norris and C. Oppenheim, “Comparing Alternatives to the Web of Science for Coverage of the Social Sciences’ Literature,” Journal of Informetrics, vol. 1, no. 2, pp. 161–169, Apr. 2007, doi: 10.1016/j.joi.2006.12.001.
[22] P. Lorenz-Spreen, L. Oswald, S. Lewandowsky, and R. Hertwig, “A systematic review of worldwide causal and correlational evidence on digital media and democracy,” Nat Hum Behav, pp. 1–28, Nov. 2022, doi: 10.1038/s41562-022-01460-1.
[23] “Leaderboard. Chatbot Arena (formerly LMSYS): Free AI Chat to Compare & Test Best AI Chatbots,” Chatbot Arena. Accessed: Feb. 28, 2025. [Online]. Available: https://huggingface.co/spaces/lmarena-ai/arena-leaderboard
[24] Y. Ye et al., “Integrating artificial intelligence with mechanistic epidemiological modeling: a scoping review of opportunities and challenges,” Nat Commun, vol. 16, no. 1, p. 581, Jan. 2025, doi: 10.1038/s41467-024-55461-x.
[25] J. Bandy, “Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits,” Proc. ACM Hum.-Comput. Interact., vol. 5, no. CSCW1, p. 74:1-74:34, Apr. 2021, doi: 10.1145/3449148.
[26] “generative AI,” Oxford English Dictionary. 2023. Accessed: Feb. 28, 2025. [Online]. Available: https://www.oed.com/dictionary/generative-ai_n
[27] A. P. Siddaway, A. M. Wood, and L. V. Hedges, “How to Do a Systematic Review: A Best Practice Guide for Conducting and Reporting Narrative Reviews, Meta-Analyses, and Meta-Syntheses,” Annu Rev Psychol, vol. 70, pp. 747–770, Jan. 2019, doi: 10.1146/annurev-psych-010418-102803.
[28] Google, “Gemma 2,” Kaggle. Accessed: Nov. 23, 2025. [Online]. Available: https://www.kaggle.com/models/google/gemma
[29] D. Macko et al., “Beyond speculation: Measuring the growing presence of LLM-generated texts in multilingual disinformation,” Mar. 29, 2025. doi: 10.1109/MC.2025.3592765.
[30] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “QLORA: efficient finetuning of quantized LLMs,” in Proceedings of the 37th International Conference on Neural Information Processing Systems, in NIPS ’23, Dec. 2023, pp. 10088–10115.
[31] J. P. T. Higgins and S. Green, Eds., Cochrane Handbook for Systematic Reviews of Interventions, 5.1.0. The Cochrane Collaboration, 2011. Accessed: Jul. 25, 2022. [Online]. Available: https://training.cochrane.org/handbook
[32] R. Ryan and Cochrane Consumers and Communication Review Group, “Heterogeneity and subgroup analyses in Cochrane Consumers and Communication Group reviews: Planning the Analysis at Protocol Stage,” Dec. 2016. [Online]. Available: http://cccrg.cochrane.org

[33] W. Viechtbauer, “Conducting meta-analyses in R with the metafor package,” Journal of Statistical Software, vol. 36, no. 3, 2010, Accessed: Dec. 05, 2025. [Online]. Available: https://doi.org/10.18637/jss.v036.i03
[34] M. Aria and C. Cuccurullo, “bibliometrix: An R-tool for comprehensive science mapping analysis,” Journal of Informetrics, vol. 11, no. 4, pp. 959–975, 2017.
[35] M. Harrer, P. Cuijpers, T. Furukawa, and D. Ebert, Doing Meta-Analysis with R: A Hands-On Guide. New York: Chapman and Hall/CRC, 2021. doi: 10.1201/9781003107347.
[36] C. Doss et al., “Deepfakes and scientific knowledge dissemination,” Sci Rep, vol. 13, no. 1, p. 13429, Aug. 2023, doi: 10.1038/s41598-023-39944-3.
[37] A. Lewis, P. Vu, R. M. Duch, and A. Chowdhury, “Deepfake detection with and without content warnings,” R. Soc. open sci., vol. 10, no. 11, p. 231214, Nov. 2023, doi: 10.1098/rsos.231214.
[38] S. Iacobucci, R. De Cicco, F. Michetti, R. Palumbo, and S. Pagliaro, “Deepfakes Unmasked: The Effects of Information Priming and Bullshit Receptivity on Deepfake Recognition and Sharing Intention,” Cyberpsychology, Behavior, and Social Networking, vol. 24, no. 3, pp. 194–202, Mar. 2021, doi: 10.1089/cyber.2020.0149.
[39] J. Ternovski, J. Kalla, and P. Aronow, “Negative Consequences of Informing Voters about Deepfakes: Evidence from Two Survey Experiments,” jots, vol. 1, no. 2, Feb. 2022, doi: 10.54501/jots.v1i2.28.
[40] J. Roozenbeek, C. S. Traberg, and S. van der Linden, “Technique-based inoculation against real-world misinformation,” Royal Society Open Science, vol. 9, no. 5, p. 211719, May 2022, doi: 10.1098/rsos.211719.
[41] C. Wittenberg, Z. Epstein, G. Péloquin-Skulski, A. J. Berinsky, and D. G. Rand, “Labeling AI-generated media online,” PNAS Nexus, vol. 4, no. 6, p. pgaf170, Jun. 2025, doi: 10.1093/pnasnexus/pgaf170.
[42] J. Lee and S. Y. Shin, “Something that They Never Said: Multimodal Disinformation and Source Vividness in Understanding the Power of AI-Enabled Deepfake News,” Media Psychology, vol. 25, no. 4, pp. 531–546, Jul. 2022, doi: 10.1080/15213269.2021.2007489.
[43] Michael Borenstein, Larry V. Hedges, Julian P. T. Higgins, Hannah R. Rothstein, Introduction to Meta-Analysis, 1st ed. John Wiley & Sons, Ltd, 2009. doi: 10.1002/9780470743386.
[44] Z. Irsova, H. Doucouliagos, T. Havranek, and T. D. Stanley, “Meta-analysis of social science research: A practitioner’s guide,” Journal of Economic Surveys, vol. 38, no. 5, pp. 1547–1566, 2024, doi: 10.1111/joes.12595.
[45] K. Bryanov and V. Vziatysheva, “Determinants of Individuals’ Belief in Fake News: A Scoping Review Determinants of Belief in Fake News,” PLoS One, vol. 16, no. 6, p. e0253717, Jun. 2021, doi: 10.1371/journal.pone.0253717.
[46] B. Zhang, S. J. Kim, and A. Scott, “Immunizing the Public Against AI-Generated Disinformation: Testing the Effects of Inoculation Mode and Issue Attitude on Inoculation Likelihood of Political Deepfakes,” Journalism & Mass Communication Quarterly, vol. 102, no. 4, pp. 1102–1134, Dec. 2025, doi: 10.1177/10776990251357949.
[47] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Hillsdale: Lawrence Erlbaum Associates, 1988.

APPENDICES

Appendix A: Methodological Details

Evidence Collection

In a few cases, when the P value was reported as $P < p$, it was assumed that $P = p$. For example, when an exact p value was available, such as $p = 0.015$, the exact value was used to calculate the standard error.

Some parameters, such as the date of data collection (experiment), were unavailable for a few publications. Where necessary, the date of the submission of the article to the journal was used as an approximation, or, if that was also unavailable, data collection was assumed to have occurred one year prior to the publication. When possible, the date of the generation of the misinformation content was used instead. For example, in one case [1], we used the date of the publication of the misinformation (December 2023 instead of November 2024), as even a year can make a significant difference in the context of rapidly developing textual LLM models. For publications featuring multiple experiments conducted on different dates that analyzed LLMs of varying capacities, we staggered the coded data collection dates by at least two months. This ensured that our meta-regression model treated more advanced models as later temporal developments when sequencing the experiments.

Assessing Heterogeneity

We assessed between-study heterogeneity for this group of studies using two measures: I^2 and τ^2. "When I^2 is near zero, the observed variability is mostly down to sampling error; when I^2 is near 100, most of the observed variability reflects differences in population effect sizes" [2]. However, the I^2 measure has disadvantages, and it is often appropriate to focus on discussing τ^2 levels, which are "insensitive to the number of studies, and their precision" [3]. A psychology-focused review found that close replication studies reported an average τ of 0.09, but studies that pooled more diverse experiments reported an average τ of between 0.31 and 0.35 [4]. In the domain of misinformation research, a high level of heterogeneity is a common issue [5].

Appendix B: Publications Reviewed using Meta-Analysis

[1] S. Barari, C. Lucas, and K. Munger, "Political deepfakes are as credible as other fake media and (sometimes) real media," The Journal of Politics, vol. 87, no. 2, pp. 510–526, Apr. 2025, doi: 10.1086/732990.

[2] S. D. Bray, S. D. Johnson, and B. Kleinberg, "Testing human ability to detect 'deepfake' images of human faces," Journal of Cyber Security, vol. 9, no. 1, p. tyad011, Jan. 2023, doi: 10.1093/cybsec/tyad011.

[3] V. Danry, P. Pataranutaporn, M. Groh, and Z. Epstein, "Deceptive explanations by large language models lead people to change their beliefs about misinformation more often than honest explanations," in Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan: ACM, Apr. 2025, pp. 1–31, doi: 10.1145/3706598.3713408.

[4] C. Doss et al., "Deepfakes and scientific knowledge dissemination," Scientific Reports, vol. 13, no. 1, p. 13429, Aug. 2023, doi: 10.1038/s41598-023-39944-3.

[5] S. S. El Mokadem, "The Effect of Media Literacy on Misinformation and Deep Fake Video Detection," Arab Media & Society, Issue 35, winter/spring 2023, Aug. 2023, doi: 10.70090/SM23EMLM.

[6] J. A. Goldstein, J. Chao, S. Grossman, A. Stamos, and M. Tomz, "How persuasive is AI-generated propaganda?," PNAS Nexus, vol. 3, no. 2, p. pgae034, Feb. 2024, doi: 10.1093/pnasnexus/pgae034.

[7] M. Hameleers, T. G. L. A. Van Der Meer, and T. Dobber, "You won't believe what they just said! The effects of political deepfakes embedded as vox populi on social media," Social Media + Society, vol. 8, no. 3, July 2022, doi: 10.1177/20563051221116346.

[8] M. Hameleers, T. G. L. A. van der Meer, and T. Dobber, "They would never say anything like this! Reasons to doubt political deepfakes," European Journal of Communication, vol. 39, no. 1, pp. 56–70, Feb. 2024, doi: 10.1177/02673231231184703.

[9] Y. Hwang and S.-H. Jeong, "Generative artificial intelligence and misinformation acceptance: An experimental test of the effect of forewarning about artificial intelligence hallucination," Cyberpsychology, Behavior, and Social Networking, Feb. 2025, doi: 10.1089/cyber.2024.0407.

[10] Y. Hwang, J. Y. Ryu, and S.-H. Jeong, "Effects of disinformation using deepfake: The protective effect of media literacy education," Cyberpsychology, Behavior, and Social Networking, vol. 24, no. 3, pp. 188–193, Mar. 2021, doi: 10.1089/cyber.2020.0174.

[11] S. Iacobucci, R. De Cicco, F. Michetti, R. Palumbo, and S. Pagliaro, "Deepfakes unmasked: The effects of information priming and bullshit receptivity on deepfake recognition and sharing intention," Cyberpsychology, Behavior, and Social Networking, vol. 24, no. 3, pp. 194–202, Mar. 2021, doi: 10.1089/cyber.2020.0149.

[12] S. Kreps, R. M. McCain, and M. Brundage, "All the news that's fit to fabricate: AI-generated text as a tool of media misinformation," Journal of Experimental Political Science, vol. 9, no. 1, 2022, doi: 10.1017/XPS.2020.37.

[13] J. Lee and M. Hameleers, "Effects of health-related deepfakes on misperceptions: Moderating effects of issue relevance and accuracy motivation," Media Psychology, pp. 1–30, Sept. 2024, doi: 10.1080/15213269.2024.2401539.

[14] J. Lee and S. Y. Shin, "Something that they never said: Multimodal disinformation and source vividness in understanding the power of AI-enabled deepfake news," Media Psychology, vol. 25, no. 4, pp. 531–546, Jul. 2022, doi: 10.1080/15213269.2021.2007489.

[15] H. Lu and H. Chu, "Let the dead talk: How deepfake resurrection narratives influence audience response in prosocial contexts," Computers in Human Behavior, vol. 145, p. 107761, Aug. 2023, doi: 10.1016/j.chb.2023.107761.

[16] H. Lu and S. Yuan, "'I know it's a deepfake': The role of AI disclaimers and comprehension in the processing of deepfake parodies," Journal of Communication, vol. 74, no. 5, pp. 359–373, Oct. 2024, doi: 10.1093/joc/jqae022.

[17] S. J. Nightingale and H. Farid, “AI-synthesized faces are indistinguishable from real faces and more trustworthy,” Proceedings of the National Academy of Sciences of the United States of America, vol. 119, no. 8, p. e2120481119, Feb. 2022, doi: 10.1073/pnas.2120481119.
[18] S. Y. Shin and J. Lee, “The effect of deepfake video on news credibility and corrective influence of cost-based knowledge about deepfakes,” Digital Journalism, vol. 10, no. 3, pp. 412–432, Apr. 2022, doi: 10.1080/21670811.2022.2026797.
[19] E. R. Spearing et al., “Countering AI-generated misinformation with pre-emptive source discreditation and debunking,” Royal Society Open Science, vol. 12, no. 6, p. 242148, June 2025, doi: 10.1098/rsos.242148.
[20] J. Ternovski, J. Kalla, and P. Aronow, “Negative consequences of informing voters about deepfakes: Evidence from two survey experiments,” Journal of Online Trust and Safety, vol. 1, no. 2, Feb. 2022, doi: 10.54501/jots.v1i2.28.
[21] C. Vaccari and A. Chadwick, “Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news,” Social Media + Society, vol. 6, no. 1, p. 2056305120903408, Jan. 2020, doi: 10.1177/2056305120903408.
[22] M. Wack, C. Ehrett, D. Linvill, and P. Warren, “Generative propaganda: Evidence of AI’s impact from a state-backed disinformation campaign,” PNAS Nexus, vol. 4, no. 4, p. pgaf083, Apr. 2025, doi: 10.1093/pnasnexus/pgaf083.
[23] T. Weikmann, J. L. Egelhofer, and S. Lecheler, “Beyond credibility: The effects of different forms of visual disinformation,” Journalism & Mass Communication Quarterly, vol. 102, no. 4, pp. 1020–1043, Dec. 2025, doi: 10.1177/10776990251357299.
[24] T. Weikmann, H. Greber, and A. Nikolaou, “After deception: How falling for a deepfake affects the way we see, hear, and experience media,” The International Journal of Press/Politics, vol. 30, no. 1, pp. 187–210, Jan. 2025, doi: 10.1177/19401612241233539.
[25] C. Wittenberg, Z. Epstein, G. Péloquin-Skulski, A. J. Berinsky, and D. G. Rand, “Labeling AI-generated media online,” PNAS Nexus, vol. 4, no. 6, p. pgaf170, June 2025, doi: 10.1093/pnasnexus/pgaf170.

Appendix C: Supplementary Tables

Table C1. Boolean Search in the Web of Science 2025 Syntax.

(((TI=(disinformation OR misinformation OR propaganda OR "fake news" OR rumo OR "misleading information" OR hallucination* OR deepfake OR "false information" OR conspirac* OR "fake content")) OR AB=(disinformation OR misinformation OR propaganda OR "fake news" OR rumo* OR "misleading information" OR hallucination* OR deepfake OR "false information" OR conspirac* OR "fake content")) OR KP=(disinformation OR misinformation OR propaganda OR "fake news" OR rumo* OR "misleading information" OR hallucination* OR deepfake OR "false information" OR conspirac* OR "fake content"))*

AND

(TI=("artificial intelligence" OR AI OR "large language model" OR "neural network*" OR "deep learning" OR "machine learning" OR gpt* OR gemini OR DeepSeek OR Step OR Grok OR Yi OR Qwen OR Claude OR Vicuna OR MPT OR LLaMA OR WizardLM OR XGen OR ChatGLM OR "generative content" OR "generative model*" OR "generative AI" OR "generative artificial" OR "transformer model*" OR "AI-powered" OR "AI powered" OR "chatbot*" OR "neural text generation" OR "text generation model")*

OR

AB=("artificial intelligence" OR AI OR "large language model" OR "neural network*" OR "deep learning" OR "machine learning" OR gpt* OR gemini OR DeepSeek OR Step OR Grok OR Yi OR Qwen OR Claude OR Vicuna OR MPT OR LLaMA OR WizardLM OR XGen OR ChatGLM OR "generative content" OR "generative model*" OR "generative AI" OR "generative artificial" OR "transformer model*" OR "AI-powered" OR "AI powered" OR "chatbot*" OR "neural text generation" OR "text generation model")*

OR

KP=("artificial intelligence" OR AI OR "large language model" OR "neural network*" OR "deep learning" OR "machine learning" OR gpt* OR gemini OR DeepSeek OR Step OR Grok OR Yi OR Qwen OR Claude OR Vicuna OR MPT OR LLaMA OR WizardLM OR XGen OR ChatGLM OR "generative content" OR "generative model*" OR "generative AI" OR "generative artificial" OR "transformer model*" OR "AI-powered" OR "AI powered" OR "chatbot*" OR "neural text generation" OR "text generation model"))*

Table C2. Codes Used to Synthesize the Research Evidence.

Code	Description of the code	Variables
Eligibility (Abstract, Title, and Keywords coding)		
English	Abstract is available in English.	0 = no 1 = yes
Misinformation	Publication discusses any aspect of "misinformation" or any related phenomenon, that is, • misleading information, • disinformation, • fake news, • propaganda, • rumors, • credibility of information, • manipulation, • hallucination • deepfakes as a key object of study. If there are several objects, "misinformation" should be at least one of no more than three of them.	0 = no 1 = yes
Generative AI	The publication focuses on Generative AI: An algorithm, a system, or a piece of software designed to produce output, especially text or visuals, such as "deepfakes," previously thought to require human intelligence, typically by using machine learning to extrapolate from large collections of data.	0 = no 1 = yes
Empirical	The study includes an experiment or analysis of quantitative or qualitative data that produces evidence-based claims. It must not be a purely opinion or position paper, although position papers with substantial empirical components are eligible.	0 = no 1 = yes
Method	Methods used to collect and analyze data. Multiple choice.	0 = other/None 1 = experimental (RCT) 2 = quasi-experimental study 3 = quantitative Survey
Measure Effects	Does the publication report on the results describing the impact, effect, or consequences of a countermeasure that aims to address GenAI-produced misinformation?	0 = no 1 = yes
Measure Subject	[If "Measure Effects" = "0", skip this Question] Is a countermeasure that aims to address Generative AI-produced misinformation focused on improving human subject practices, experiences, perceptions, or similar (e.g., "literacy about deepfakes") or a computational model (e.g., "RAG for GenAI models")?	0 = no 1 = computational model 2 = human subjects 3 = model and human

People Effect	Does the publication report on the impact, effect, or consequences of exposure to GenAI-produced misinformation on human subjects?	0 = no 1 = yes
Full Paper Coding		
GenAI Model	What GenAI model does this study design focus on?	0 = none 1 = GPT OpenAI models 2 = Bard/Gemini 2 = Mistral 3 = Grok 4 = Qwen 5 = Claude 6 = Vicuna 9 = MPT 10 = LLaMA 11 = WizardLM 12 = XGen 13 = ChatGLM 14 = Faceswap 15 = unspecified textual model 100 = other, custom, or not specified visual model
GenAI Model Specify	Please specify the model.	[open coding]
Geographic Context	Where were the participants, if any, recruited from?	1 = USA 2 = Canada 3 = EU 4 = UK 5 = East Asia (e.g., China) 6 = South Asia (e.g., India) 7 = Oceania 8 = Africa 9 = Central/South America 10 = MTurk, Prolific, or similar 12 = Other 100 = Not specified
Date of Data Collection	Specify the month and year when the data collection was finished, or, if that information is not available, when the content for the experiment was generated. If the data collection date is not specified, use the start date of the data collection.	[open coding]
Misinformation Domain	What is the domain of misinformation that the study focuses on in relation to GenAI use?	0 = not specified or mixed

		1 = health 2 = political 3 = gender 4 = other 5 = climate
Format Type	What is the format of GenAI-produced misinformation content that is the focus of the study?	0 = unspecified 1 = text 2 = text and image 3 = audiovisual passive (video, television, movie, YouTube clip) 4 = audio only 5 = images only 6 = other
People Type	Who are the human participants who were recruited for the study?	1 = college students 2 = patients 4 = mixed community and college sample 5 = other convenience sample 6 = MTurk or similar convenience sample firms 100 = not specified
People Type Specify	Specify what exact human participants were recruited.	[open coding]
Measures Proposed	Countermeasures mitigating the impact of misinformation. The proposed measures should mitigate the impact of misinformation generated with GenAI. Multiple choice.	NA = nothing proposed 0 = broad 1 = advertisement policy 2 = labeling or watermarks 3 = content or account moderation 4 = content reporting 5 = content distribution & sharing 6 = corrective information materials 7 = disinformation disclosure 8 = information literacy and education 9 = redirection 10 = security or verification 11 = AI-assisted or automatic fact-checking infrastructure 12 = AI detection systems

		model improvements (e.g., more data, changing parameters) 13 = Human-only fact-checking 14 = Other computational model additions 15 = human-led improvements (e.g., improving prompts, humans giving feedback to a model via GUI) 100 = other
Measures Proposed Specify	Briefly describe specific countermeasures proposed to address the problem of misinformation.	[open coding]
Stimulus and Procedure	Specify the details of the experiment.	[open coding]

Source: IPIE.
Note: Codes are ordered as they appear in the codebook provided to coders. For replication purposes, a version with examples is available upon request. Codes were partially adapted from SR2023.1.

Table C3. LLM Prompt.

You are an expert research assistant and social scientist specializing in human-computer interaction, media psychology, communication, sociology, and the societal impacts of Artificial Intelligence. Your primary objective is to act as a screener for a meta-analysis. You will be provided with a list of academic publications, each with a title and an abstract. Your task is to meticulously review each publication and determine its eligibility for inclusion based on a strict criterion: Research Methodology. You must evaluate each publication independently. criterion: Research Methodology The publication MUST report on a study using an Experimental Design, Randomized Controlled Trial (RCT), or a Quasi-Experimental Study (QES). Refer to these precise definitions: Experimental Design / Randomized Controlled Trial (RCT): The study involves an intervention where participants are randomly assigned to different conditions. This includes within-subject designs where participants are exposed to multiple conditions in a randomized order, as well as natural experiments where a naturally occurring event creates random-like assignment. Quasi-Experimental Study (QES): The study involves an intervention or comparison between groups but lacks random assignment. Groups are assigned to a countermeasure or condition based on self-selection (e.g., users choosing to use a tool) or administrator selection (e.g., a class of students receiving an educational module while another does not). It must involve a comparison between these non-equivalent groups or a pre-test/post-test design around an intervention. What to EXCLUDE based on methodology: Purely theoretical or conceptual papers. Literature reviews or systematic reviews. Opinion pieces or perspective articles that do not present new empirical data from an experiment. Technical papers describing only a system's architecture or performance (e.g., the accuracy of a detection model) without testing its effect on human subjects. Paper relying only on synthetic data: papers that do not study humans, only systems, algorithms, content, etc. Purely correlational or descriptive survey studies that do not involve a manipulated intervention or a clear comparison between non-equivalent groups exposed to different conditions. In output, print out NNs and reasoning only for those articles that satisfy the criterion and are included. Do not mention not-included articles in the output.

Table C4. Publications by Type of Evidence.

Authors Year	Title	Misinformation Effect	Corrective Information	Labeling
Barari, Lucas and Munger 2025	Political deepfakes are as credible as other fake media and (sometimes) real media		✓	✓
Bray et al. 2023	Testing human ability to detect "deepfake" images of human faces		✓	
Danry, Pataranutaporn, Groh and Epstein 2025	Deceptive explanations by large language models lead people	✓		
Doss et al. 2023	Deepfakes and scientific knowledge dissemination			
Goldstein et al. 2024	How persuasive is AI-generated propaganda?	✓		
Hameleers, Van Der Meer and Dobber 2022	You won't believe what they just said! The effects of political deepfakes embedded as vox populi on social media	✓		
Hameleers, Van Der Meer and Dobber 2024	They would never say anything like this! Reasons to doubt political deepfakes	✓		
Hwang and Jeong 2025	Generative artificial intelligence and misinformation acceptance: An experimental test of the effect of forewarning about artificial intelligence hallucination			
Hwang, Ryu and Jeong 2021	Effects of disinformation using deepfake: The protective effect of media literacy education	✓	✓	
Iacobucci 2021	Deepfakes unmasked: The effects of information priming and bullshit receptivity on deepfake recognition and sharing intention		✓	
Kreps, McCain and Brundage 2022	All the news that's fit to fabricate: AI-generated text as a tool of media misinformation	✓		
Lee and Hameleers 2024	Effects of health-related deepfakes on misperceptions: Moderating effects of issue relevance and accuracy motivation	✓		
Lee and Shin 2022	Something that they never said: Multimodal disinformation and source vividness in understanding the power of AI-enabled deepfake news	✓		✓
Lu and Chu 2023	Let the dead talk: How deepfake resurrection narratives influence audience response in prosocial contexts			✓
Lu and Yuan 2024	"I know it's a deepfake": The role of AI disclaimers and comprehension	✓	✓	

	in the processing of deepfake parodies			
Nightingale and Farid 2022	AI-synthesized faces are indistinguishable from real faces and more trustworthy	✓		
Shin and Lee 2022	The effect of deepfake video on news credibility and corrective influence of cost-based knowledge about deepfakes	✓	✓	
Spearing et al. 2025	Countering AI-generated misinformation with preemptive source discreditation and debunking	✓	✓	✓
Ternovski, Kalla and Aronow 2022	The negative consequences of informing voters about deepfakes: Evidence from two survey experiments		✓	
Vaccari and Chadwick 2020	Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news		✓	
Wack, Ehrett, Linvill and Warren 2025	Generative propaganda: Evidence of AI's impact from a state-backed disinformation campaign	✓		
Weikmann, Egelhofer and Lecheler 2025	Beyond credibility: The effects of different forms of visual disinformation	✓		
Weikmann, Greber and Nikolaou 2025	After deception: How falling for a deepfake affects the way we see, hear, and experience media	✓		
Wittenberg et al. 2025	Labeling AI-generated media online		✓	

Appendix D: Appendices References

[1] M. Wack, C. Ehrett, D. Linvill, and P. Warren, 'Generative propaganda: Evidence of AI's impact from a state-backed disinformation campaign', PNAS Nexus, vol. 4, no. 4, p. pgaf083, Apr. 2025, doi: 10.1093/pnasnexus/pgaf083.

[2] J. P. T. Higgins and S. G. Thompson, 'Quantifying Heterogeneity in a Meta-Analysis', Statistics in Medicine, vol. 21, no. 11, pp. 1539–1558, 2002, doi: 10.1002/sim.1186.

[3] M. Harrer, P. Cuijpers, T. Furukawa, and D. Ebert, Doing Meta-Analysis with R: A Hands-On Guide. New York: Chapman and Hall/CRC, 2021. doi: 10.1201/9781003107347.

[4] A. H. Linden and J. Hönekopp, 'Heterogeneity of Research Results: A New Perspective From Which to Assess and Promote Progress in Psychological Science', Perspect Psychol Sci, vol. 16, no. 2, pp. 358–376, Mar. 2021, doi: 10.1177/1745691620964193.

[5] K. Bryanov and V. Vziatysheva, 'Determinants of Individuals' Belief in Fake News: A Scoping Review Determinants of Belief in Fake News', PLoS One, vol. 16, no. 6, p. e0253717, Jun. 2021, doi: 10.1371/journal.pone.0253717.

[6] J. A. C. Sterne et al., 'RoB 2: a revised tool for assessing risk of bias in randomised trials', BMJ, p. l4898, Aug. 2019, doi: 10.1136/bmj.l4898.

[7] L. G. Smithers, A. C. P. Sawyer, C. R. Chittleborough, N. M. Davies, G. Davey Smith, and J. W. Lynch, 'A systematic review and meta-analysis of effects of early life non-cognitive skills on academic, psychosocial, cognitive and health outcomes', Nat Hum Behav, vol. 2, no. 11, Art. no. 11, Nov. 2018, doi: 10.1038/s41562-018-0461-x.

[8] J. P. T. Higgins et al., 'The Cochrane Collaboration's Tool for Assessing Risk of Bias in Randomised Trials', BMJ, vol. 343, p. d5928, Oct. 2011, doi: 10.1136/bmj.d5928.

[9] T. d. Stanley and H. Doucouliagos, 'Picture This: A Simple Graph That Reveals Much Ado About Research', Journal of Economic Surveys, vol. 24, no. 1, pp. 170–191, 2010, doi: 10.1111/j.1467-6419.2009.00593.x.

[10] W. Viechtbauer, 'Selection Models — selmodel', Github. Accessed: Dec. 05, 2025. [Online]. Available: https://wviechtb.github.io/metafor/reference/selmodel.html#details-1

ACKNOWLEDGMENTS

Contributors

Drafting authors: Aliaksandr Herasimenka (Consulting Scientist, United Kingdom), Sebastián Valenzuela (IPIE Chief Science Officer and Chair of the Science & Methodology Committee, Chile), Shelley Boulianne (IPIE Science & Methodology Committee Member, Canada), Frank Esser (IPIE Science & Methodology Committee Member, Switzerland), Lisa M. Given (IPIE Science & Methodology Committee Member, Canada/Australia), Stephan Lewandowsky (IPIE Science & Methodology Committee Member, Australia/United Kingdom), Eva M. Navarro-López (IPIE Science & Methodology Committee Member, Spain/UK/Mexico), Philip Howard (IPIE President and CEO, Canada/UK). Research Assistants: Anna George and Xianlingchen Wang. Independent General Reviews: George Georgarakis and Mathias Harrer. Design: Domenico Di Donna. Copyediting: Beverley Sykes. We gratefully acknowledge support from the IPIE Secretariat: Lola Gimferrer, Jessica Gold, Wiktoria Schulz, Donna Seymour, Anna Staender, and Alex Young.

Funders

The International Panel on the Information Environment (IPIE) gratefully acknowledges the support of its funders. For a full list of funding partners please visit www.ipie.info. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the IPIE and do not necessarily reflect the views of the funders.

Declaration of Interests

IPIE reports are developed and reviewed by a global network of research affiliates and consulting scientists who constitute focused Scientific Panels and contributor teams. All contributors and reviewers complete declarations of interests, which are reviewed by the IPIE at the appropriate stages of work.

Preferred Citation

An IPIE *Summary for Policymakers* provides a high-level precis of the state of knowledge and is written for a broad audience. An IPIE *Synthesis Report* makes use of scientific meta-analysis techniques, systematic review, and other tools for evidence aggregation, knowledge generalization, and scientific consensus building, and is written for an expert audience. An IPIE *Technical Paper* addresses particular questions of methodology, or provides a policy analysis on a focused regulatory problem. All reports are available on the IPIE website (www.IPIE.info).

This document should be cited as:

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.)], “Confronting Misinformation Produced with Generative AI: A Meta-Analysis of Experimental Scientific Evidence,” Zurich, Switzerland: IPIE, 2026. Synthesis Report, SR2026.2, doi: 10.61452/UGTR3022.

Authors' AI Contribution Statement

Large Language Models (LLMs), specifically Copilot and Gemma 2, were used to support specific stages of this work, including generating R code for data cleaning, transformation, and visualization, and for screening the abstracts, keywords, and titles of scientific publications. All LLM-assisted or LLM-generated content was critically reviewed, verified, and revised by at least one human contributor, who takes full responsibility for the final version of the work.

Copyright Information

ABOUT THE IPIE

The International Panel on the Information Environment (IPIE) is an independent and global science organization committed to providing the most actionable scientific knowledge about threats to the world's information environment. Based in Switzerland, the mission of the IPIE is to provide policymakers, industry, and civil society with independent scientific assessments on the global information environment by organizing, evaluating, and elevating research, with the broad aim of improving the global information environment. Hundreds of researchers from around the world contribute to the IPIE's reports.

For more information, please contact the International Panel on the Information Environment (IPIE), secretariat@IPIE.info. Seefeldstrasse 123, P.O. Box, 8034 Zurich, Switzerland.

International Panel on
the Information
Environment

Seefeldstrasse 123
P.O. Box 8034 Zurich
Switzerland

ISBN 978-3-03983-015-2