The Role of AI in the Disinformation Landscape

Project	
Project number:	*101158277*
Project name:	*BENEDMO: Flemish-Dutch collaboration to monitor, study and fight the spread of disinformation in the Dutch-language region*
Project acronym:	*BENEDMO*

Deliverable	
Deliverable number:	*D4.3*
Deliverable name:	*White paper demonstrating the contribution of emerging technologies to the effects of disinformation*
Name Author:	*Teresa Weikmann et al.*
Affiliation:	*University of Amsterdam*
Name reviewer:	*Guy De Pauw & Brahim Zarouali*
Affiliation:	*Textgain & University of Leuven*

Co-funded by the European Union

Disclaimer

Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.

Inhoudsopgave

Co-funded by the European Union

The Role of AI in the Disinformation Landscape – Insights from the BENEDMO Research Lab and Beyond

Teresa Weikmann, Michael Hameleers, Marina Tulin & Claes de Vreese

Executive summary:
Concerns about the widespread application of Artificial Intelligence (AI) in the generation of deceptive information and its amplification on social media are widespread. At the same time, however, we currently lack systematic insights into its prevalence, effects, and counterstrategies (Farooq & de Vreese, 2026). Hence, although the alarm on deepfakes and AI-powered disinformation rings loudly in policy, media, and social discourse (WEF, 2025), current studies on the effects of AI-disinformation do not suggest that deepfakes are more powerful or deceptive than other modes of disinformation (Barari et al., 2021; Weikmann et al., 2025a). Yet, the scope of existing research is limited or preliminary, and may fail to keep up with the fast pace in which generative AI is developing and democratizing. Even more so, mapping short-term effects after exposure may not fully grasp the wider consequences that generative AI has for society, organizations, and the status of truth and authenticity more generally. Drawing on the efforts and outcomes of the **Dutch-Flemish hub of the Digital Media Observatory (BENEDMO)**, we lay out the consequences of AI in disinformation and showcase evidence in the effectiveness of counter-responses, such as the labelling of generative AI on social media, and fact-checking interventions. We further offer a practical perspective by including the views of practitioners that professionally deal with disinformation and verification: What kind of challenges do they identify regarding the role of AI in the disinformation ecosystem? What are their needs to counter or prevent negative consequences of disinformation in a changing technological context?

This white paper proceeds in the following ways. **First**, we r**eview the state-of-the-art of research on the prevalence and effects** of AI-generated and amplified disinformation. **Second**, drawing on **original findings from the BENEDMO research lab**, we critically discuss the evidence for the effectiveness of counterstrategies and preventative measures that shield society from the potential harms of AI-generated disinformation, including

Co-funded by the European Union

from our own research. **Third**, we offer an **expert perspective on the role of AI in verification ecosystems**, including the potential barriers experienced by practitioners in analyzing and refuting AI-generated disinformation. Together, these three steps aim to offer a comprehensive overview of how AI has evolved the disinformation ecosystem, and a forward-looking perspective on potential trends and developments that we should consider in the rapid enrollment and embedding of AI in the processing, consumption, and generation of digital information flows.

What is the role of AI in disinformation?

We adopt a wide definition of AI-powered disinformation. In line with von Sikorski and Hameleers (2025), we regard the influence of AI on disinformation across different levels, including the technological advancements used to generate, manipulate, or fabricate synthetic media, and the ways in which disinformation is amplified or disseminated through different online sources. We therefore adopt the following definition throughout this paper:

> *AI-powered disinformation concerns the application of AI in the generation, alteration, dissemination, or manipulation of intentionally false information, which includes deceptive visual content (i.e., videos and images), speech, or textual information.*

Such disinformation is created or disseminated to advance political, financial, personal, ideological, or any other strategic aims of the creator or disseminator. AI-powered disinformation can come in various shapes and forms: It can, among other things, include the generation of synthetic videos to impersonate known political figures (Hameleers et al., 2026), the creation of fake news articles (Tulin et al., 2025), the generation of false images for social media posts (Weikmann et al., n.d. a), use of bots and trolls to push inauthentic content (e.g., Starbird et al., 2023), or manipulated audio messages used in the context of corporate fraud. Here, we would like to stress that the field on gen-AI is moving at an extremely rapid pace, and that academic research is therefore not always up to date. In line with this, the evidence of academic papers presented here is often based on work that is under review and not yet published. Therefore, some conclusions we draw are preliminary and (as the playing field of AI advances) in need of constant updating.

Co-funded by the European Union

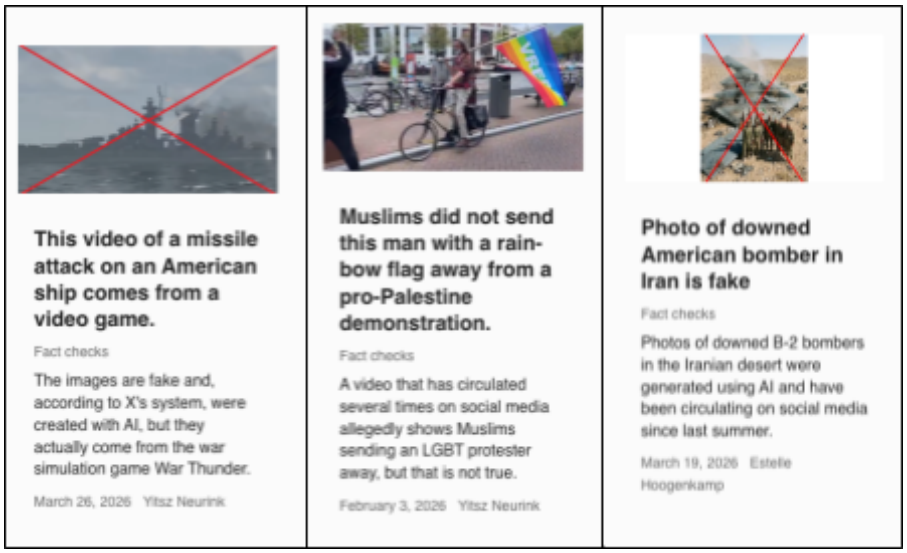

Figure 1. *Recent articles by Dutch fact-checkers Nieuwscheckers on the use of AI*

To date, academic research on the share of AI-generated content in the disinformation landscape is scarce. The limited evidence available suggests that – at least in the early 2020s – AI-generated content is scarce (Brennen et al., 2021; Hameleers, 2025; Weikmann & Lecheler, 2023; Yang et al., 2023). Visual disinformation at this stage mainly consists of decontextualized visuals in which authentic footage is taken from a different context, location, or timeframe to alter its suggested interpretation (Hameleers, 2025). Yet, it is crucial to stress that the literature may not be up to date with the latest advancements in generative AI, which have democratized the use of AI in deceptive contexts, allowing anyone with an internet connection to generate and disseminate deceptive content through social media. The embedding of Grok, Claude, ChatGPT, and other related applications in people's everyday communication tools have made it easier and more acceptable to use AI in social media posting. Although this may not always be done by powerful actors with the intention to cause harm, it can also be used by citizens in interpersonal communication (i.e., group chats in WhatsApp) to attack or harm targeted individuals. Revenge porn and other forms of cyber-bullying can be made increasingly more credible and harmful with the use of AI. In a political context, authentic opinion leaders, social media influencers, and bots and trolls may rely on AI to generate deceptive political content to attack the opposed camp, or to legitimize counter-factual claims on reality. Although the use of bots and trolls has been well-documented in the context of disinformation and propaganda (e.g., Starbird et al., 2023; DiResta & Goldstein, 2025), AI offers unprecedented options to automize the generation of deceptive content by fake and inauthentic social media accounts. Thus, beyond troll factories in which real paid workers are hired to push false political narratives, the process of information influence campaigns can be upscaled and amplified by automating the process of undermining content.

Co-funded by the European Union

In the Dutch context, the CampAIgn Tracker (Votta & Kruschinski, 2025) offers a leading example of an attempt to capture the use of generative AI for political goals. The latest tracker is based on all posts on selected social media (i.e., Facebook, Instagram, TikTok, and X) from political parties and candidates who ran for the Dutch parliamentary elections 2025, as well as selected influencers and commentators. Results show an increasingly higher prevalence of AI-generated images in the month leading up to the Dutch parliamentary elections, specifically by far-right political actors. In terms of content, the AI-generated images appear to speak to existing animosities around migration and the loss of Dutch traditions, which was the main campaign topic of far-right political parties in the last election. Prominent AI-generated images depicting Geert Wilders (leader of the largest far-right party) seemed to be circulated with the goal of speaking to the imagination of a strong leader who will shield the people from the threats of migration. In the context of this example, it should be noted that the line between deceptive or harmful information versus disinformation is thin and complex. Hence, although the use of AI in campaign messages involves the (often undisclosed) use of generative AI to create synthetic media, the messages themselves are not always pushing clearly false factual claims. Often, they reflect negative campaigning and satire to mock the opponent or emphasize the strength of the in-group. This also has ramifications for interventions: Fact-checking the statements themselves may not be worthwhile, whereas emphasizing the use of AI and the strategies behind political campaigning can be a more viable path.

Figure 2. *AI imagery monitored via Votta & Kruschinski's (2025) CampAIgnTracker*

Co-funded by the European Union

The increased presence of AI-generated content, especially during election times, is also mirrored in citizens' perceptions of the prevalence of disinformation. Farooq et al. (2025) found that in the context of the most recent European parliament elections, citizens in the Netherlands, Germany and Poland reported to be concerned about the misuse of generative AI. They indicated to encounter AI-generated content monthly, especially leading up to the elections and right after, and they estimated that almost 30% of the AI-generated content they see can be classified as disinformation, meaning that it is was deliberately deceiving. Yet, again, it can be questioned whether such content encountered in election times would fit stricter definitions stressing false content and inaccurate factual claims as being central to disinformation.

Effect studies on AI-generated disinformation have mostly focused on the effects of deepfakes compared to other modes of disinformation (e.g., Barari et al., 2025; Hameleers et al., 2026; Weikmann et al., 2025a). These studies consistently indicate that deepfake videos are not substantially more credible or persuasive compared to other forms of disinformation (Barari et al., 2025). Yet, in some cases, deepfakes have stronger effects on the delegitimization of political actors or worldviews (e.g., Dobber et al., 2020). Yet, it should be emphasized that the data collection of various recent research projects is based on AI that is already outdated and did not take into account the full potential that new tools of generative AI bring to the table. Especially the democratization of tools embedded on social media which are freely available to users (i.e., Grok, ChatGPT) can increase the use of generative AI in disinformation, and have longer-term consequences on people's perceptions of authenticity and legitimacy. If anyone can use AI tools to generate realistic but fake media from scratch, this may impact the value and status of (visual) evidence. Hence, beyond the direct impact of deepfakes on measured credibility or authenticity, the presence of generative AI on social media may result in the declined trust in authentic visuals, or increased relativism toward the value of evidence in general (Tulin et al., 2025; Weikmann et al., 2025b). People may have come to accept that AI tools are widespread and easy to use for everyone. When they come across visual content that causes suspicion in some shape or form (i.e., because of glitches or deviations from existing beliefs), they may be likely to associate it with AI. This also works the other way around: In times of increased epistemic uncertainty and factual relativism, AI-generated content that does not cause suspicion as it reassures existing beliefs may be deemed authentic.

Overall, observational evidence and the work of fact-checkers and investigative journalism shows that **the use of AI is a new reality in the disinformation ecosystem,** and a growing challenge for the experts working on it (Wouters et al., 2026). In high-stakes election periods, we observe that disinformation created with generative AI is a new reality (e.g., deCheckers, 2026; Nieuwscheckers, 2025), whereas academic research has not yet

Co-funded by the European Union

sufficiently studied these dynamics, and the effects that the democratization of AI has on longer term trust in evidence, and the overall status of facts, truths, and visual proof.

Here, we also note the harms of so-called abliterated LLMs. These are versions of existing open-source models that have their guard rails removed and will no longer refuse to generate specific types of content. Since these models open-source, they can be run locally, further removing any type of accountability. When refusal behavior is removed from LLMs, they may generate unethical, hate-driven, and illegal content at the disposal of malicious communicators.

The effectiveness of counterstrategies

In this section, we review existing research and knowledge on the effectiveness of correcting AI-powered disinformation. We generally discern between two different approaches: (1) pre-bunking literacy initiatives that aim to make audiences more resilient to disinformation by teaching them how to detect AI-generated content; (2) fact-checking or debunking interventions that flag AI in disinformation contexts, including policy or regulatory approaches such as labelling or watermarking.

Pre-bunking literacy interventions

With regards to this first route of instilling resilience before people encounter AI-powered disinformation, there is only a limited amount of empirical evidence on how literacy interventions may be applied to the new context of AI. Outside the context of AI, media literacy interventions are found to be successful in helping people to detect deception and misinformation (Hameleers, 2022; Lu et al., 2024; Li et al., 2025). However, existing work has mostly focused on text-based mis- or disinformation, while AI-powered disinformation and deepfakes that contain seemingly realistic visual disinformation have received more scholarly attention only recently (e.g., von Sikorski & Hameleers, 2025). Yet, despite the visual and AI-driven reality of contemporary disinformation settings, systematic studies that test the real-life effectiveness of interventions in the context of AI-powered disinformation are relatively scarce.

In support of the potential of pre-bunking literacy interventions in this AI-driven context, Huang and Hu (2025) show that different media literacy interventions presented before showing deepfake videos enhanced people's ability to detect deepfakes. These interventions also improved self-perceived AI literacy. Hwang et al. (2021) also show that including a literacy intervention reduced the negative impact of exposure to deepfake videos. Deng and Ahmed (2025) report findings that are equally optimistic: Based on survey data, they conclude that higher levels of media literacy can shield people from engaging with deepfakes. In contrast, Tulin et al. (2025) found that an AI literacy

intervention that teaches about the capabilities of generative AI for the creation of disinformation (e.g., fake news articles) resulted in mixed effects. While it increased citizens' ability to successfully detect fake news articles as fake, it did not improve their overall discernment, such that they also incorrectly classified genuine news articles as fake. In optional responses to an open question about what respondents thought of the intervention, the majority reported a version of feeling "scared", that they do not know "what to believe anymore" and that "anything can be fake."

This brings us to an even more substantial point that is also raised in the general literature on inoculation or pre-bunking messages. Inoculation has been coined as an effective intervention for preventing the harms of mis- and disinformation as it exposes people to a small dose of the threatening information (i.e., a short deepfake fragment) after which people are offered the tools to recognize and counter such content themselves (Roozenbeek & van der Linden, 2019). Although these interventions are found to be effective in reducing the harm of disinformation by increasing detection skills, they may also come at a cost: They emphasize the risk of being deceived online and offer only a limited and non-exhaustive set of suggestions that may adequately point people to some forms of disinformation, whereas they may not help people to detect more nuanced and complex forms of deception. Even more so, and potentially more worrisome, people may also wrongly apply their critical mindset to information that resembles disinformation in some respects, whilst being honest and accurate information (also see Hoes et al., 2024). As case in point, people were first taught that a wrong number of fingers or a clock that would always be set at the same time would suggest "fingerprints" of AI. Yet, later iterations of AI developments have fixed these glitches, which may mean that people look for indicators that are no longer part of more advanced AI applications. At the same time, when an authentic video or image shows a clock at the same time or has irregularities around the hands of a depicted person due to perspective or irregularities in the image, people may wrongly associate such content with AI-powered disinformation.

De-bunking, fact-checking and labelling

Next to warning people up front, AI-powered disinformation can also be corrected and refuted after its spread. In general, research on fact-checking interventions has been hopeful: Exposure to corrective information that factually refutes false statements of disinformation lowers the credibility of falsehoods across the board (Walter et al., 2020). Yet, there are also barriers in place that may prevent people from selecting or accepting corrective information – which mostly relates to confirmation biases (e.g., Hameleers & van der Meer, 2020; Thorson, 2016). In other words, again situated in high choice information ecologies, people may in real life avoid fact-checking information, or only selectively attend to corrective information that reassures their existing beliefs.

Co-funded by the European Union

In the field of AI-powered disinformation, the task of correcting disinformation becomes even more complicated. The ease and speed with which such content can be created enables an increasing variety in depictions and narratives, especially when it comes to visuals. This trend further complicates mitigation efforts and creates an especially demanding environment for professionals seeking to correct and fact-check hyper-realistic falsehoods. As content becomes more varied, it is no longer sufficient to debunk individual false images or claims; interventions must instead be able to broadly reduce the perceived credibility of AI-generated visual disinformation across a wide range of content and contexts. This becomes especially complicated in a social media environment, where AI-generated images regularly find traction. Here, mitigating efforts often constitute one-size-fits-all solutions, such as watermarking AI-generated content (Weikmann et al., n.d. a).

In a series of recent experiments, the BENEDMO lab tested the effectiveness of currently available platform interventions against AI-powered disinformation, hereby focusing on the Dutch context. Specifically, our goal was to understand whether content moderation efforts such as AI-labels/watermarks, fact-check labels or community notes would be able to reduce the overall credibility of AI-powered disinformation posts, or lower the extent to which Dutch citizens would believe in the false claim portrayed. Overall, our results suggest that interventions are largely ineffective. We tested these labels across a variety of false claims, either pertaining to disinformation about immigration or the consequences of climate change in the Netherlands, which were all considered moderately credible regardless of which label was applied. However, community notes achieved a reduction in belief in false claims about immigration, when isolating responses to this topic, suggesting the promise of citizen-based approaches. Yet, in a second experiment (Weikmann et al., n.d. b), we find that direct interventions by fact-checkers, that is, fact-checkers directly commenting under posts to correct disinformation (Opgenhaffen, 2025), showed overall the strongest effects for reducing belief in a false claim. We found this again in an immigration disinformation context, this time focusing on the issue of allocation of social housing in the Netherlands. In this study, we also found that it may be fruitful to not only factually correct false claims, but also address potential intent behind it, as participants who had strong anti-migrant attitudes found the post less credible when receiving such a contextual intervention (Weikmann et al., n.d. b).

Overall, our findings underscore that there is no simple solution to combating AI-powered disinformation, and that one-size-fits-all interventions are unlikely to be effective. This is also in line with the perceptions of EDMO experts who work on mitigating AI-driven disinformation. Our survey amongst professionals reveals that they believe that more than one intervention is needed (Weikmann et al., 2026b). Fact-checkers, for instance, highlight

Co-funded by the European Union

that platforms are for an important part responsible, while researchers highlight the crucial role of media literacy. Moreover, the extent to which certain measures reduce citizens' belief in false or misleading content depends not only on the specific claim being presented, but also on individuals' pre-existing beliefs and attitudes toward the issue in question. Simultaneously, platform-based interventions face important practical limitations. For instance, citizen-based approaches such as community notes rely on lay participation, which may be error-prone, biased, and susceptible to misinformation. They also often depend on referenced fact-checking articles, and should not be considered a replacement of traditional fact-checking, but rather a complementary tool – even though Meta has famously ended their partnerships with fact-checking organizations in the US. Moreover, AI-labels are not consistently applied, and do not address potential disinformation in practice – they only indicate whether content has been AI-generated or not. Ultimately, platforms bear much of the responsibility for intervening and labelling, but opaque practices and hard-to-evaluate mitigation measures make it difficult to hold them accountable (Weikmann et al., 2026a).

Given the partially hopeful but also mixed findings on different approaches, we should note that much of the existing evidence is restricted to artificial survey contexts and mainly based on immediate effects of exposure to interventions. The reality of the disinformation landscape is substantially more complicated. In high-choice information environments, people may selectively avoid literacy or corrective messages: These may not stand out as compared to the false information it intends to correct. Another substantial barrier is that experimental evidence and other studies illustrate a demand effect: People first see an intervention that tells them what to do to comply and then must fulfil a related task that tests these instructions. Therefore, the positive effect of interventions found in studies may not be replicated in real life settings where there is no clear instruction about a set task that is connected to the intervention.

Insights from the BENEDMO consortium

The BENEDMO network brings together a diverse group of disinformation experts who work on different facets of the problem and face growing challenges in their daily practice (Wouters et al., 2026). In this third part of the white paper, we present expert perspectives from the BENEDMO consortium on the role of AI in the verification ecosystem. Overall, their observations and assessments broadly align with the conclusions we draw from academic research.

Co-funded by
the European Union

Media literacy expert Zara Mommerency (Mediawijs) focuses on the societal and educational dimensions of AI. She notes, in line with academic research, a growing general distrust of information that is reinforced by the rise of AI. Citizens' responses, she observes, are often both curious and resistant, as many lack the skills to fully understand AI's risks and opportunities. To address this, Mommerency calls for sustained investment in critical media and AI literacy across all ages and sectors, so that citizens not only know how to use AI tools, but can also question, interpret, and evaluate AI-mediated information. As AI becomes ever more embedded in people's everyday lives, work, and communication, learning to engage with it critically and responsibly will be essential. Initiatives such as the upcoming AI Act will ideally provide clearer safeguards and accountability, helping people feel that they are not left alone to navigate the impact of AI on their own.

> *"We see both curiosity and resistance towards AI, often because people are not yet fully aware of its opportunities and risks. Media and AI literacy are therefore essential: people need the knowledge and skills to critically and confidently engage with these (new) technologies, without being left behind by rapid technological change. At the same time, we notice a growing distrust towards online information in general. As AI-generated content becomes harder to distinguish from reliable sources, some people become skeptical not only of misleading content, but also of trustworthy journalism and verified information, which leads to cynicism and confusion."* - **Zara Mommerency**

> *"AI affects our work both in terms of AI-generated disinformation and in terms of editorial work. On the one hand, we are seeing an increase in AI-generated disinformation. On the other hand, more tools that utilize AI are becoming available, and there is also the broader question of how AI is being integrated into general editorial operations and how media brands position themselves in a rapidly changing environment."* - **Luc van Bakel**

Moreover, AI is reshaping the **media sector**, as underscored by Luc van Bakel (Editor-in-Chief, VRT NWS). He highlights how AI directly affects newsroom practice, in part due to the rapid proliferation of AI-generated disinformation, which is also visible in platform monitoring tools such as the CampAIgnTracker. At the same time, there are still very few reliable instruments for detecting AI-generated content: Google SynthID is one of the only tools currently available, yet its capabilities remain limited – a challenge for journalists that also emerged in the BENEDMO expert survey (Weikmann et al., 2025). From this vantage point, the most urgent needs are structural: a stronger regulatory framework for large social media platforms, more reliable and scalable tools for detecting AI-generated content, and faster systems for monitoring what circulates online.

Co-funded by the European Union

Naturally, AI-powered disinformation also reshapes **disinformation detection** and **fact-checking** practices. Guy De Pauw, CEO and co-founder of artificial-intelligence company Textgain, approaches disinformation detection through computational text analysis. He argues that AI is necessary to cope with the sheer volume of problematic content spreading online through the possibilities of generative AI. He also emphasizes that a human in the loop remains crucial and that there is no fully automated "fact-checking machine". Similarly, digital journalism scholar and member of factcheck.vlaanderen Michaël Opgenhaffen sees AI as an opportunity to make fact-checking more accessible, adaptable, and effective. While he acknowledges that AI presents both opportunities and challenges for journalism, he highlights its potential to help fact-checkers translate content, tailor it to different audiences, and distribute it across multiple platforms, thereby expanding the reach of verified information. At the same time, he emphasizes the importance of maintaining established verification standards and remaining flexible in response to rapid technological change. Rather than focusing on predicting future developments, he argues that the priority should be building resilient institutions and professionals capable of evolving alongside AI.

> *"We believe AI itself can (and should be) deployed to handle the massive scale at which bad actors can now operate. We simply cannot afford not to develop AI to counteract what's coming at us in the years ahead."* – **Guy De Pauw**

> *"One area where I see significant potential is the use of AI to make fact-checking more accessible and impactful. AI can help translate fact-checks into different languages, adapt them for different audiences, and repackage them for various platforms and formats. This allows fact-checkers to reach more people with reliable information without compromising the accuracy of the original verification work."*
> – **Michaël Opgenhaffen**

Conclusion

Building on the insights from state-of-the-art academic research, our own findings from the BENEDMO hub, and professional insights from the BENEDMO consortium, we conclude that AI has not fundamentally altered the underlying logic of disinformation, but it has transformed its **scale, speed, and accessibility**, and in doing so it reshapes how citizens, institutions, and information systems relate to evidence and truth. Current research shows that deepfakes and AI-generated content are not consistently more persuasive than traditional forms of disinformation, yet the rapid democratization of generative tools, embedded in everyday platforms and communication practices, means that deceptive content can be produced and disseminated more easily, more cheaply, and by a wider range of actors. Monitoring efforts such as the CampAIgn Tracker and the experiences of fact-checkers already signal a growing presence of AI-generated images and narratives in high-stakes contexts such as elections, particularly around polarizing topics. The most consequential impact of AI may therefore be less about single pieces of content and more

about **epistemic consequences**: the erosion of trust in authentic visual evidence. Counterstrategies offer only partial relief. Pre-bunking and literacy interventions can increase people's ability to detect AI-generated content and bolster self-perceived AI literacy, but they also risk reinforcing over-skepticism and information nihilism. Platform-based measures – e.g., AI labels, fact-check flags or community notes – show limited and often issue-dependent effects on credibility and belief in false claims, as shown by experimental evidence from the BENEDMO lab. Meanwhile, journalists and verification professionals operate in a structurally unfavourable environment: tools for reliably detecting AI-generated content remain immature, access to platform data is restricted, and regulatory frameworks are still catching up. Overall, the evidence suggests that there is no single technical or communicative fix. Effective mitigation will require **layered, adaptive strategies** that combine platform accountability and regulation, investment in robust and up-to-date verification capacities, and sustained, age-spanning media and AI literacy that fosters critical engagement without drifting into cynicism. At the same time, research must close the gap with practice by systematically tracking the prevalence and evolution of AI-powered disinformation across platforms, examining long-term effects on trust and democratic resilience, and evaluating interventions in realistic, high-choice media environments. In sum, AI-driven disinformation should be treated not as a temporary anomaly, but as a structural condition of contemporary information ecologies that demands continuous monitoring, experimentation, and institutional learning.

References

Barari, S., Lucas, C., & Munger, K. (2025). Political Deepfakes Are as Credible as Other Fake Media and (Sometimes) Real Media. *The Journal of Politics, 87*(2), 510–526. https://doi.org/10.1086/732990

Brennen, J. S., Simon, F. M., & Nielsen, R. K. (2021). Beyond (Mis)Representation: Visuals in COVID-19 Misinformation. *International Journal of Press/Politics*, *26*(1). https://doi.org/10.1177/1940161220964780

deCheckers. (2026, May 7). *Factcheck: Deze AI-afbeelding toont niet de humanitaire hulpvloot naar Gaza*. deCheckers. https://decheckers.be/factchecks/factcheck-deze-ai-afbeelding-toont-niet-de-humanitaire-hulpvloot-naar-gaza

Deng, R., & Ahmed, S. (2025). Perceptions and paradigms: An analysis of AI framing in trending social media news. *Technology in Society*, *81*, 102858. https://doi.org/10.1016/j.techsoc.2025.102858

DiResta, R., & Goldstein, J. A. (2025). Full-Spectrum Propaganda in the Social Media Era. *Security Studies*, *34*(4), 714–750. https://doi.org/10.1080/09636412.2025.2563254

Co-funded by the European Union

Dufour, N., Pathak, A., Samangouei, P., Hariri, N., Deshetti, S., Dudfield, A., Guess, C., Escayola, P. H., Tran, B., Babakar, M., & Bregler, C. (2024). *AMMeBa: A Large-Scale Survey and Dataset of Media-Based Misinformation In-The-Wild*.

Farooq, A., & de Vreese, C. (2026). Disinformation and its Sociopolitical Context. In *Disinformation: A Multi-Disciplinary Analysis* (pp. 3-22). Cham: Springer Nature Switzerland.

Farooq, A., van den Hoogen, E., Tulin, M., & de Vreese, C. (2025). Generative AI Generating Concerns: Citizens' Perspectives During the 2024 European Elections. *The International Journal of Press/Politics*, 19401612251376060.

Hameleers, M. (2022). Separating truth from lies: Comparing the effects of news media literacy interventions and fact-checkers in response to political misinformation in the US and Netherlands. *Information, communication & society, 25*(1), 110-126.

Hameleers, M. (2025). The visual nature of information warfare: the construction of partisan claims on truth and evidence in the context of wars in Ukraine and Israel/Palestine. *Journal of Communication*, *75*(2), 90–100. https://doi.org/10.1093/joc/jqae045

Hameleers, M., & van der Meer, T. G. L. A. (2020). Misinformation and Polarization in a High-Choice Media Environment: How Effective Are Political Fact-Checkers? *Communication Research*, *47*(2), 227–250. https://doi.org/10.1177/0093650218819671

Hameleers, M., van der Meer, T. G. L. A., Tulin, M., & Dobber, T. (2026). Radical Right-Wing Political Deepfakes Can Successfully Delegitimize Targeted Political Actors: Evidence From Three-wave Experiments in the US and The Netherlands. *Communication Research*. https://doi.org/10.1177/00936502261421437

Hoes, E., Aitken, B., Zhang, J., Gackowski, T., & Wojcieszak, M. (2024). Prominent misinformation interventions reduce misperceptions but increase scepticism. *Nature Human Behaviour*, *8*(8), 1545–1553. https://doi.org/10.1038/s41562-024-01884-x

Huang, G., & Hu, B. (2025). "A Warning is Not Enough. Teach Me How to Spot Deepfakes.": Testing Media Literacy Interventions for Combating Deepfakes. *Science Communication*, 10755470251382889.

Hwang, Y., Ryu, J. Y., & Jeong, S.-H. (2021). Effects of Disinformation Using Deepfake: The Protective Effect of Media Literacy Education. *Cyberpsychology, Behavior, and Social Networking*. https://doi.org/10.1089/cyber.2020.0174?casa_token=i2AK7QOuIrcAAAAA%3AfiMQbmbGry4gPcNNBs8DJATkIyIhPiXY9IerQ4uv4Qpowo

Li, C., Zhang, M., & Xie, K. (2025). Enhancing digital literacy in children and adolescents: a meta-analysis of school-based interventions. *Information, Communication & Society*, 1-28.

Co-funded by the European Union

Lu, C., Hu, B., Bao, M. M., Wang, C., Bi, C., & Ju, X. D. (2024). Can media literacy intervention improve fake news credibility assessment? A meta-analysis. *Cyberpsychology, Behavior, and Social Networking, 27*(4), 240-252.

Nieuwscheckers. (2025, November 18). *Image Whisperer helpt bij beeldverificatie*. Nieuwscheckers. https://nieuwscheckers.nl/image-whisperer-helpt-bij-beeldverificatie/

Opgenhaffen, M. (2025). Studying Fact-Checks on Social Media: Using a Real-Life Fact-Checking Platform to Explore the Feasibility and Practice of Direct Content Interventions. In *Research Methods for Social Media Journalism* (pp. 52-64). Routledge.

Roozenbeek, J., & van der Linden, S. (2019). The fake news game: actively inoculating against the risk of misinformation. *Journal of Risk Research*, *22*(5), 570–580. https://doi.org/10.1080/13669877.2018.1443491

Starbird, K., DiResta, R., & DeButts, M. (2023). Influence and improvisation: Participatory disinformation during the 2020 US election. Social Media+ Society, 9(2), 20563051231177943.

Thorson, E. (2016). Belief Echoes: The Persistent Effects of Corrected Misinformation. *Political Communication*, *33*(3), 460–480. https://doi.org/10.1080/10584609.2015.1102187

Tulin, M., Pantazi, M., Starke, C., Sivolap, M., & Dobber, T. (2025). Generative AI and Disinformation| Echoes of Doubt: Exposure to Information About Generative AI Decreases Believability of News. *International Journal of Communication, 19*, 24-24.

von Sikorski, C., & Hameleers, M. (2025). Disinformation in the Age of Artificial Intelligence (AI): Implications for Journalism and Mass Communication. *Journalism & Mass Communication Quarterly*, *102*(4), 941–957. https://doi.org/10.1177/10776990251375097

Votta, F., & Kruschinski, S. (2025). *CampAIgn Tracker*. https://www.campaigntracker.nl/info.html

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-Checking: A Meta-Analysis of What Works and for Whom. *Political Communication*, *37*(3), 350–375. https://doi.org/10.1080/10584609.2019.1668894

WEF. (2025). Global Risks Report 2025: Conflict, Environment and Disinformation Top Threats. https://www.weforum.org/press/2025/01/global-risks-report-2025-conflict-environment-and-disinformation-top-threats/

Weikmann, T. E., Tulin, M., Hameleers, M., & de Vreese, C. (n.d. b). *Checking Facts or Adding Context? Evaluating the Effectiveness of Direct Content Interventions Against AI-driven Disinformation on Social Media*

Co-funded by the European Union

Weikmann, T. E., Tulin, M., Hameleers, M., & de Vreese, C. (n.d. a). *No easy fix to countering AI-generated visual disinformation: The (in)effectiveness of AI-labels, fact-check labels and community notes*. https://doi.org/10.31219/osf.io/8237p_v1

Weikmann, T., & Lecheler, S. (2023). Cutting through the Hype: Understanding the Implications of Deepfakes for the Fact-Checking Actor-Network. *Digital Journalism*, 1–18. https://doi.org/10.1080/21670811.2023.2194665

Weikmann, T., De Pauw, G., Tulin, M., Hameleers, M. & de Vreese, C. (2026a). Monitoring the Implementation of the Code of Conduct on Disinformation through Structural Indicators: Gaps, Evidence, and Policy Implications. *BENEDMO.*

Weikmann, T., Egelhofer, J. L., & Lecheler, S. (2025a). Beyond Credibility: The Effects of Different Forms of Visual Disinformation. *Journalism & Mass Communication Quarterly*, *102*(4), 1020–1043. https://doi.org/10.1177/10776990251357299

Weikmann, T., Greber, H., & Nikolaou, A. (2025b). After Deception: How Falling for a Deepfake Affects the Way We See, Hear, and Experience Media. *The International Journal of Press/Politics*, *30*(1), 187–210. https://doi.org/10.1177/19401612241233539

Weikmann, T., Wouters, F., Tulin, M., Hameleers, M., de Vreese, C., Zarouali, B., & Opgenhaffen, M. (2026b). On the same page? Experts are mostly, but not always aligned about disinformation in times of generative AI. *Harvard Kennedy School Misinformation Review*, *7*(2). https://doi.org/10.37016/mr-2020-196

Wouters, F., Weikmann, T., Zarouali, B., Opgenhaffen, M., Hameleers, M., Tulin, M., & de Vreese, C. (2026). The Burden of Truth? Risks of Countering Disinformation in Flanders and the Netherlands. *Tijdschrift Voor Communicatiewetenschap*, *54*(2), 198–220. https://doi.org/10.5117/TCW2026.2.004.WOUT

Yang, Y., Davis, T., & Hindman, M. (2023). Visual misinformation on Facebook. *Journal of Communication*. https://doi.org/10.1093/joc/jqac051

Co-funded by the European Union