



## UvA-DARE (Digital Academic Repository)

### No easy fix to countering AI-generated visual disinformation: The (in)effectiveness of AI-labels, fact-check labels and community notes

Weikmann, T.E.; Tulin, M.; Hameleers, M.; de Vreese, C.H.

**DOI**

[10.31219/osf.io/8237p\\_v1](https://doi.org/10.31219/osf.io/8237p_v1)

**Publication date**

2025

**Document Version**

Submitted manuscript

**License**

CC BY

[Link to publication](#)

**Citation for published version (APA):**

Weikmann, T. E., Tulin, M., Hameleers, M., & de Vreese, C. H. (2025). *No easy fix to countering AI-generated visual disinformation: The (in)effectiveness of AI-labels, fact-check labels and community notes*. OSF Preprints. [https://doi.org/10.31219/osf.io/8237p\\_v1](https://doi.org/10.31219/osf.io/8237p_v1)

**General rights**

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

**Disclaimer/Complaints regulations**

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

***No easy fix to countering AI-generated visual disinformation:  
The (in)effectiveness of AI-labels, fact-check labels and community notes***

T. E. Weikmann<sup>1</sup>, M. Tulin<sup>1</sup>, M. Hameleers<sup>1</sup> and C. H. de Vreese<sup>1</sup>

<sup>1</sup> Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Amsterdam, The Netherlands

Contact corresponding author: t.e.weikmann@uva.nl

**Abstract**

As generative AI makes it easier to create synthetic visuals, AI-driven visual disinformation is becoming more common on social media. However, while much research highlights its potential harm, less is known about how to reduce its potential to mislead. In this study, we therefore conducted a preregistered online experiment in the Netherlands (N=1,018) to test the effectiveness of various platform interventions: (1) AI labels or “watermarks,” (2) fact-check labels, and (3) community notes. We tested how effective these sources are in lowering credibility of the false visual and belief in the false claim it portrays across two polarizing topics: climate change and immigration. Overall, the interventions showed no significant differences in effectiveness. This was the case when pooling both topics together and for climate-change related disinformation in isolation. However, for visual disinformation about immigration, community notes were most effective, especially among participants with strong anti-migrant views. Our findings suggest that while labeling has limited impact overall, its effectiveness varies by context, and no one-size-fits-all solution exists for combating AI-generated visual disinformation.

**Keywords:** AI label, community notes, disinformation intervention, fact-checking, generative AI, social media, visual disinformation

**Introduction**

The rise of generative artificial intelligence (AI) has introduced powerful tools for creating synthetic visual content, which can be used to construct and spread disinformation on social media (Peng et al., 2024). Such AI-generated visual disinformation spins false narratives across a variety of polarizing topics, such as international conflicts, political actors or environmental issues (Hameleers, 2024). Visual disinformation has been shown to be perceived as authentic information (Hameleers et al., 2022; Shen et al., 2021), negatively affect attitudes towards politicians (Dobber et al., 2021), lead to a loss of trust in visuals (Weikmann et al., 2024) and news on social media

(Vaccari & Chadwick, 2020). However, while these consequences are relatively well-researched, effective strategies to *mitigate* its impact remain less understood (Dan, 2025), unlike for the widely studied interventions targeting text-based disinformation (e.g., Hameleers & van der Meer, 2020; Jia et al., 2022; Martel & Rand, 2023; Tulin et al., 2024; Walter et al., 2020). Specifically, it remains unclear which social media platform interventions – where much of AI-generated visual disinformation can be found – effectively reduce perceptions of credibility for manipulated images and the adoption of false beliefs aligned with their content. This is important, as visuals may be more persuasive than textual falsehoods, as they can offer visible proof of events that never occurred (e.g., Weikmann & Lecheler, 2023b). Furthermore, even less is known about how these effects may vary across different types of visual disinformation and issue contexts, as most existing studies have focused on single topics (e.g., Jia et al., 2022; Kreps & Kriner, 2022) and single stimuli (e.g., Koch et al., 2023).

In recent years, research on disinformation – defined as deliberately created and disseminated false information (Chadwick & Stanyer, 2022) – has increasingly focused on visual formats, particularly amid the rise of so-called “deepfakes” (Dan et al., 2021). While deepfakes typically refer to real video footage altered using AI techniques, advances in generative AI also allow for the creation of entirely synthetic visuals from text prompts. AI-generated still images are easier to produce than traditional deepfakes, as they require no technical expertise such as machine learning skills (Sordo et al., 2025). Although the exact prevalence of AI-generated (visual) disinformation remains uncertain (Simon et al., 2023), there is consensus that its production has become more accessible and of better quality (Matich et al., 2025; Wittenberg et al., 2024). Moreover, research suggests that visual disinformation spans across a variety of different topics (Hameleers, 2024), usually providing evidence for false claims made in accompanying text, for instance within a social media post (Brennen et al., 2021; Yang et al., 2023). Notable recent examples include fabricated images of wildfires in Los Angeles showing unaffected areas engulfed in flames, and fake images depicting Donald Trump being arrested.

Very large online platforms provide some of the most prominent tools to freely create and distribute AI-generated images (Sordo et al., 2025). For instance, X’s embedded tool Grok has been noted for generating highly realistic content with minimal prompt restrictions, alongside tools like Google’s Gemini and Meta AI. Paradoxically, these platforms are also one of the key actors that politicians hold responsible for preventing the misuse of such tools in spreading

disinformation (De Blasio & Selva, 2021). In response, three types of interventions have emerged: AI-generated content labels (or watermarks), traditional fact-check labels, and community notes – user-generated annotations that provide context (Jia et al., 2022). Research suggests that such platform interventions are generally effective in countering disinformation on platforms (Drolsbach et al., 2024; Epstein et al., 2023; Martel & Rand, 2023). However, only little is known about the extent to which they work for visual disinformation, especially when it comes to comparing different label sources. This is important, as trust in these intervening entities may vary (e.g., Lee, 2024; Primig, 2024).

To address these gaps in the literature, we conducted a pre-registered<sup>1</sup> and pilot-tested survey-embedded experiment among a representative sample of Dutch participants (N = 1,018) comparing the sources of currently existing platform interventions in response to AI-generated disinformation, namely (a) AI-generated labels, (b) fact-checking labels and (c) user-based community notes vs. (d) a control group without intervention. We test the effects of these sources on perceived credibility of the visual and belief in the portrayed false claim. In addition, we investigate whether such effects are moderated by pre-existing attitudes towards the source as well as regarding the topics covered in AI-generated visual disinformation. Here, we focus on topics that may resonate with either left-wing or right-wing issue positions, namely climate change and immigration. Instead of using single stimuli to represent these topics, we created a variety of visual disinformation representing diverse false claims. As such, our study advances current understanding of the effectiveness of current platform interventions addressing AI-generated visual disinformation across a variety of originally created AI-generated images, issue contexts and beyond the US-context. The empirical insights may be used to inform the development of more effective platform policies in the future.

### **Visual disinformation in times of generative AI**

Visual falsehoods are nothing new, but technological advancements in fabricating misleading images have in recent years sparked increased concerns about the emergence of visual disinformation, that is, images that have been actively and deliberately manipulated to display an event that did not occur (Weikmann & Lecheler, 2023b, p. 3700). Visual disinformation differs from textual falsehoods not only in terms of its anticipated psychological impact, but first and foremost in its production (Weikmann & Lecheler, 2023b). Malicious actors now have access to a

wide array of tools and techniques to construct deceptive visual narratives, which vary in technological sophistication. Creating visual disinformation can be as easy as taking an image out of context through a false caption, or as complex as training a machine learning model to produce a deepfake video (Dan et al., 2021). Initially, sophisticated forms like deepfakes appeared less common than more basic decontextualized visual disinformation due to the technical difficulty involved in creating them (Brennen et al., 2021; Weikmann & Lecheler, 2023a).

However, technology is advancing fast, and novel text-to-image generators raise increased concerns about the potential for producing convincing visual falsehoods. For instance, when the improved model MidJourney V was launched in March 2023, a spike of AI-generated visual content was detected on X, including multiple instances of visual disinformation (Corsi et al., 2024). Later that year, Google launched its generative AI model Gemini, followed by Elon Musk introducing Grok on X (Sordo et al., 2025). With these advancements in place, AI-generated visuals were increasingly found on X during the 2024 U.S. Presidential Election, raising concerns about their misuse as disinformation targeting politicians (Chen et al., 2025).

What makes this development particularly challenging is the ease and speed with which tools like Gemini and Grok allow malicious actors, and even ordinary citizens, to creatively fabricate and illustrate disinformation on virtually any topic. While visual disinformation was initially found to be rather repetitive (Yang et al., 2023), recent evidence suggests that we are dealing with an increasing variety of visuals (Dufour et al., 2024). This further complicates mitigation efforts. Specifically, as content becomes more diverse, it is no longer enough to debunk individual false images or claims. Instead, effective interventions must be capable of broadly reducing the credibility of AI-generated visual disinformation across diverse content and contexts. Building on this, our study tests the effectiveness of interventions through different labelling sources across a variety of AI-generated visual disinformation, both in terms of image and claim content.

### **The effectiveness of platform interventions on mitigating AI-generated disinformation**

At present, social media platforms employ two primary interventions to mitigate false or misleading content after their spread: fact-checking labels and community notes. In a European context, this aligns with the Code of Practice on Disinformation, under which platforms have committed to strengthening fact-checking efforts and making fact-checks more visible (European Commission, 2022). Additionally, Community Notes enable platform users to contribute context,

clarifications, and potential corrections beneath potentially misleading posts. This crowdsourced content moderation system is already widely implemented on X. In January 2025, Meta announced its decision to discontinue fact-checking in the U.S. and shift entirely to a community notes model. When it comes to AI-generated content, the European Union’s Artificial Intelligence Act mandates social media platforms to label AI-generated material (European Parliament, 2023). Broadly speaking, this means there are three key entities that could potentially label and rectify AI-generated visual disinformation on social media: AI-generated labels, independent fact-checkers, and social media users via community notes. Previous research indicates that these different label sources may differ in their effectiveness (e.g., Jia et al., 2022).

### *Reducing credibility of the false visual and belief in the false claim*

Regarding the successful mitigation of AI-generated visual disinformation, we consider two outcomes as essential: (1) reducing the credibility of the false visual and (2) correcting the misinformed belief in the false claim it portrays. The first outcome refers to the image itself; in theory, visuals may be perceived as believable by default. As described in Sundar’s MAIN-model (modality, agency, interactivity, navigability), their modality cues the realism heuristic – a mental shortcut that leads individuals to relate them to the real world (Sundar, 2008). Research suggests that this effect may be driven by the quality of the false image: the more lifelike it appears, the more credible it is perceived (Dobber et al., 2021; Weikmann et al., 2024). A successful label intervention would therefore make clear that this visual is not in fact real, thus lowering its perceived credibility. The second outcome relates to the claim supported by the image. People may be more likely to believe a false claim when it is accompanied by an image that appears to provide supporting evidence. Specifically, a visual may make the false memory of an event more vivid, thus enhancing its persuasiveness (Weikmann & Lecheler, 2023b). This outcome depends more on the narrative and context than on the image’s realism (Lee et al., 2023; Lee & Hameleers, 2024). Thus, an effective intervention should prevent individuals from accepting a false claim even when paired with seeming visual proof. Although these outcomes are assessed independently, we anticipate them to occur concurrently, as both involve *reducing* the impact of AI-generated visual disinformation.

To effectively counter disinformation, interventions must overcome a range of cognitive and affective barriers (Ecker et al., 2022). Nevertheless, research shows that corrections can reduce

credibility of disinformation and/or belief in false claims. In the case of interventions through labels, two key mechanisms may be at play. First, a label may alter how disinformation is processed in the moment of encountering it, prompting individuals to re-evaluate it rather than accepting it as ‘truth by default’ (Pennycook et al., 2018). Second, if the label comes from a credible source, it can enhance the perceived reliability of the correction, making it more trustworthy than the disinformation itself (Prike & Ecker, 2023).

In line with this, prior studies generally find that labels are effective tools for countering disinformation on social media (Martel & Rand, 2023), although their effects tend to be modest (Clayton et al., 2020). Wording and design are crucial elements in this case. Labels must state that the information is ‘false’ or ‘misleading’ (Epstein et al., 2023; Wittenberg et al., 2024) – and the more explicitly this is emphasized, the more effective (Clayton et al., 2020). Effectiveness also increases with elaborateness: labels that provide contextual information and state that the content has been fact-checked tend to outperform simple tags (Drolsbach et al., 2024). The nature of the labeled content further influences effectiveness. While text-based disinformation has been widely studied, visual content poses additional challenges, as it requires labels that are large and prominently displayed to be noticed (Oeldorf-Hirsch et al., 2020). However, only recently have scholars begun examining how best to label AI-generated visuals, leaving many open questions about best practices (Kreps & Kriner, 2022; Wittenberg et al., 2024).

Building on these insights, we will design the interventions in our experiment to be equally effective in principle, differing only in the source providing the label. Specifically, each label will specify the entity identifying the visual as AI-generated and include a brief corrective statement stating that the post is misleading, prominently displayed in a social media post. In line with the discussed research, we pose the following hypotheses for our outcome variables:

**H1a:** The presence of a labeling intervention through either an (a) AI source, (b) fact-checking source or (c) community note source will be more effective in lowering the credibility of the visual and **H1b:** agreement with the false claim compared to no intervention.

Zooming in on differences between labelling sources, the findings of earlier studies are mixed. Jia et al. (2022) show that algorithmic labels, fact-checks and community notes are all equally effective when correcting false textual information on social media, if users are liberal. Numerous studies find fact-check labels work particularly well (Koch et al., 2023), which is even

the case for people who have low trust in fact-checking (Martel & Rand, 2024). In addition, a long line of literature demonstrates the overall effectiveness of fact-checks in lowering the credibility of mis- and disinformation (Walter et al., 2020), recently also regarding deepfakes (Dan, 2025). For AI-labels, research suggests that process-based labels – that mainly have the function of declaring *how* an image was constructed – are less effective in reducing the credibility of falsehoods (Wittenberg et al., 2024). At the same time, recent studies suggest that labeling journalistic content as AI-generated can increase skepticism about veracity even when the information is accurate (Altay & Gilardi, 2024). Research on community notes is still limited, but they also appear to be effective in making mis- and disinformation less credible (Jia et al., 2022). For instance, Drolsbach et al. (2024) show that they may even increase trust in fact-checking, whereas this effect is likely driven by their elaborateness in contrast to a simple flag. Given that fact-checks appear to be effective overall, especially for visuals (Dan, 2025; Wittenberg et al., 2024) and in contrast to an AI-label (Epstein et al., 2023; Wittenberg et al., 2024), we hypothesize:

**H2a:** A fact-checking source will be more effective in lowering credibility of the visual and

**H2b:** lowering agreement with the false claim compared to an AI-source.

Lastly, because research on community notes is still limited, we ask:

**RQ1:** Is there a difference in perceived credibility of the visual between a labeling intervention through an AI source vs. a community note source or **RQ1b:** between a fact-checking source vs. a community note source?

**RQ2:** Is there a difference in agreement with the labeled false claim between a labeling intervention through an AI source vs. a community note source or **RQ2b:** between a fact-checking source vs. a community note source?

### *The moderating role of pre-existing attitudes towards the source*

The effectiveness of different labeling sources may not only depend on the source itself but also on people's baseline trust or prior attitudes toward that source. In particular, what matters for belief correction is to what extent corrective sources are considered credible (Hovland & Weiss, 1952; Metzger et al., 2003). For instance, while fact-checking can be a powerful corrective tool, its impact is often shaped by whether individuals perceive fact-checkers as credible (Martel & Rand, 2024). Some view them as representatives of elite journalistic institutions, which can reduce the effectiveness of fact-checks among those who are distrustful of mainstream media (Primig,

2024) and make them particularly effective for those with high levels of trust. In contrast, crowd-sourced community notes may be more persuasive for individuals disillusioned with traditional news outlets, as they are not seen as authoritative in the same way, and rather governed by like-minded sources (Drolsbach et al., 2024). This aligns with the belief in the “wisdom of crowds,” where people tend to trust collective judgment (Allen et al., 2021; Sundar, 2008). Supporting this, research shows that individuals are more likely to accept crowd-sourced fact-checks when they activate the ‘bandwagon’ heuristic – the tendency to believe that if many people agree on something, it is likely to be true (Banas et al., 2022). Conversely, automated labels generated by algorithms may be viewed as especially objective, a perception rooted in research on human–machine communication (Lee, 2024). Specifically, if individuals are of the opinion that machines are “objective” and “free from ideological bias” (Sundar, 2008, p. 83), a correction through an AI source may be more successful for these individuals. In line with this, we hypothesize:

**H4:** An AI source will be more effective in **a)** lowering credibility of the visual and **b)** lowering agreement with the false claim for respondents with higher levels of belief in AI.

**H5:** A fact-checking source will be more effective in **a)** lowering credibility of the visual and **b)** lowering agreement with the false claim for respondents with higher levels of trust in fact-checking.

**H6:** A community note source will be more effective in **a)** lowering credibility of the visual and **b)** lowering agreement with the false claim for respondents with higher levels of the bandwagon heuristic.

### *The role of topics portrayed in disinformation*

While research on the effectiveness of different labeling interventions has addressed a range of topics – such as climate change (Koch et al., 2023), COVID-19 (Jia et al., 2022), and national or international affairs (Wittenberg et al., 2024) – relatively little is known about how effectiveness compares across different topics. This gap is important to address, as topic-specific dynamics, such as motivated reasoning, may influence whether individuals accept corrections. Individuals rarely evaluate information objectively but are more likely to accept information that already aligns with their beliefs and are more critical if it contradicts them (Taber & Lodge, 2006). In line with this, people may be more willing to revise their beliefs when the false information already contradicts their prior attitudes (Vegetti & Mancosu, 2020). Supporting this, a meta-analysis by Walter et al.

(2020) finds that fact-checking has a stronger effect on belief revision when the correction aligns with individuals' existing views (pro-attitudinal), as opposed to when it challenges them (counter-attitudinal). Moreover, research shows that partisanship can shape the effectiveness of labeling interventions. Jia et al. (2022) compared an algorithmic label – similar to an AI-label – with a fact-check and a community note, and found that for liberal users, all three types of labels reduced the perceived believability of false posts about COVID-19, regardless of the posts' ideological leaning. For conservatives, the effectiveness of the labels varied depending on whether the misinformation aligned with their political views. If it did, the algorithmic label performed particularly well. However, it remains unclear to what extent these findings translate to AI-generated *visuals*, or how they hold across different topical domains.

To study this, we choose two prominent topics associated with disinformation resonating with opposite ends of the political spectrum: climate change and immigration. Both issues are politicized and vulnerable to disinformation (e.g., Gustafson et al., 2020; Hameleers et al., 2021) and may resonate with different ideological beliefs, therefore offering a different setting for potential confirmation biases and other patterns in responses to fact-checking. We explicitly varied the partisan emphasis across issues. Specifically, immigration was framed as a right-wing anti-immigration narrative, whereas climate change was framed from a left-wing angle that (over)emphasized the risks of anthropogenic climate change. As such, beyond studies that have mainly looked at right-leaning disinformation (see Walter et al., 2020), our study is one of the first that explicitly compares the effects of (corrected) disinformation framed with opposite ideological perspectives (but see Jia et al., 2022). Although we do not have a-priori expectations about the differences in effectiveness of corrections across these issues, we compare two contrasting issues as a robustness check and explore to what extent the effects of corrective information in response to visual disinformation are similar or different across prominent disinformation topics on left-wing versus right-wing positions. We pose the following research questions:

**RQ3:** To what extent do the underlying topics of AI-generated visual disinformation (climate change vs. immigration) affect the perceived credibility of the visual and agreement with the false claim?

**RQ4:** To what extent is this moderated by pre-existing attitudes towards either (a) climate change or (b) immigration?

## Method

We conducted a pre-registered online experiment (N = 1,018) with a between-subjects design with four experimental conditions. Participants were exposed to AI-generated visual disinformation embedded in a simulated social media post accompanied by one of the following: (a) an AI-generated label, (b) a fact-checking label, (c) a user-based community note, or (d) no label (control condition).

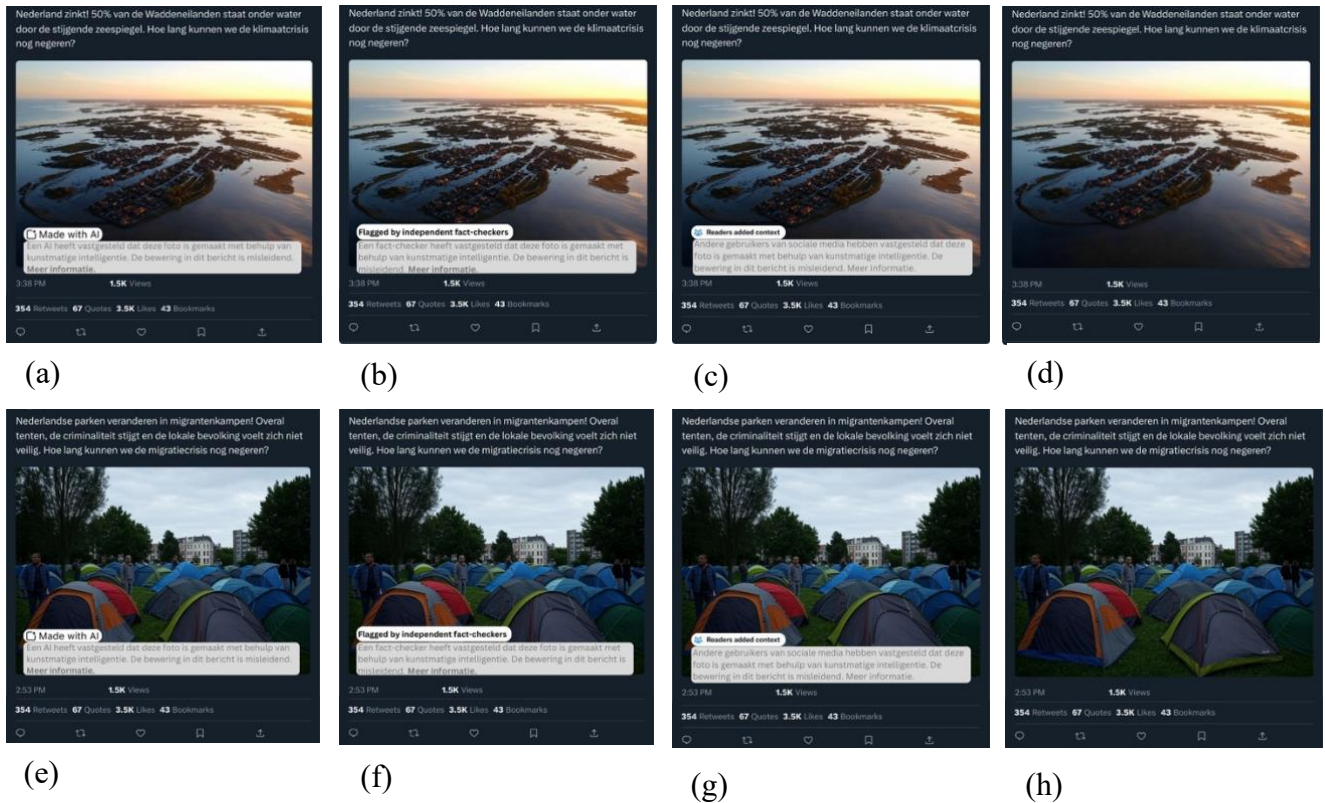
## *Stimuli*

***Creation of AI-generated visual disinformation.*** To create the visual disinformation stimuli, we used Grok, the text-to-image generator offered by X, due to its ability to produce high-quality, photorealistic outputs. Specifically, we used textual prompts asking Grok to visualize the false claims about climate change and immigration we came up with. For instance, one image depicted the Frisian Islands submerged under water, accompanied by a caption making exaggerated claims about the climate crisis – even though the islands have never been submerged. An example from the immigration topic showed AI-generated refugees pitching tents in a public park, with a caption expressing public outrage, even though such an event never occurred (see Figure 1).

***Pretest.*** After generating an initial pool of 30 images, we conducted a pretest using a non-representative sample of Dutch participants on Prolific (N = 100) to assess perceived credibility of the visuals, plausibility of the false narrative, and recognizability as AI-generated. Based on this pretest, we performed a k-means cluster analysis to group the images according to these perceptual variables. Images that were either too easily identified as fake or seen as highly convincing were excluded, to avoid floor or ceiling effects. Details of this pretest are reported in the supplementary material A.

***Label design.*** We modeled the design of the labels closely on current interventions used on major social media platforms, such as Meta and X (e.g., Instagram Help Center, 2024; Meta, 2019). While real-world implementations differ slightly in terms of placement and wording, we standardized these elements across conditions to ensure comparability. The tag itself – declaring the source of the intervention – was taken directly from Meta and X in case of the AI-label (“*Made with AI*”) and community note (“*Readers added context*”). Each label included a brief corrective message: “*A [label source] has determined that this photo was created using artificial intelligence. The claim in this post is misleading. More information.*”

The final stimulus set consisted of 10 AI-generated images with variations in narrative: five depicting exaggerated consequences of human-induced climate change and five showing misleading narratives around immigration policy. Each image was paired with one of the four labeling conditions (AI-label, fact-check, community note, or control), resulting in a total of 40 experimental stimuli (5 climate change  $\times$  4 labels + 5 immigration  $\times$  4 labels = 40), collapsed into four groups. Examples for both topics are displayed in Figure 1.



**Figure 1.** Example stimuli for topics climate change (a-d) and immigration (e-h) and for AI-label (a, e), fact-check label (b, f), community note (c, g) and control group (d, h).

### Procedure

Upon providing informed consent, participants first completed a set of demographic questions. This was followed by a series of pre-treatment measures, including: belief in the AI-heuristic, baseline trust in fact-checking, endorsement of the bandwagon heuristic (i.e., the belief that large groups are effective at distinguishing true from false information), and issue attitudes toward climate change and immigration (see ‘Measures’ below). To minimize priming effects,

participants then answered a short set of distractor questions related to their general news consumption. Participants were subsequently exposed to one of the experimental stimuli, which remained on screen for a minimum of 12 seconds (this was deemed the minimum exposure time to process the message and its content). Following exposure, we included a manipulation check asking what topic the post covered, which also functioned as a distractor. Next, we measured the dependent variables, namely perceived credibility of the visual and agreement with the false claim. A manipulation check then assessed whether participants correctly recalled the label source or recognized the absence of a label in the control condition. At the end, participants received a thorough debriefing that explained the purpose of the experiment and clarified that the visual content they had seen was AI-generated and did not depict real events. Additionally, we provided resources on how to identify visual disinformation and included a final confirmation step in which participants acknowledged their understanding that the images were not real. This study received ethical approval by the University of Amsterdam (FMG-13362).

### ***Measures***

All variables were measured using 7-point Likert scales (1 = strongly disagree, 7 = strongly agree), unless explicitly stated otherwise. The exact item wording, scale indices, and Dutch and English translations are available in the pre-registration document on the Open Science Framework (OSF)<sup>1</sup>.

### ***Pre-treatment variables***

***Belief in AI heuristic*** was measured with four items, asking participants whether they thought that “artificial intelligence (AI)” was “objective,” “neutral,” “accurate” and “fair”, following research by Lee et al., (2022) ( $M = 3.99$ ,  $SD = 1.34$ ,  $\alpha = .83$ ).

***Trust in fact-checking*** was measured with four items in line with prior work (Martel & Rand, 2024; Tsfaty & Cappella, 2003) by asking participants how they perceive fact-checkers, using examples relevant to the Dutch media context (e.g., Nieuwscheckers, deCheckers, and Knack). In particular, we wanted to know whether they thought fact-checkers were “fair,” “told the whole story,” or “could be trusted” ( $M = 4.03$ ,  $SD = 1.10$ ,  $\alpha = .79$ ).

***The bandwagon heuristic***, i.e., whether participants believed in the wisdom of the crowd was measured using three items according to Banas et al., (2022). For instance, we asked

participants whether they thought that “If many other people believe that information is true, it is probably true” ( $M = 3.59$ ,  $SD = 1.33$ ,  $\alpha = .80$ ).

**Prior attitudes towards immigration** were measured similarly to Hameleers et al. (2021), using five items that asked whether participants thought that “refugees posed a threat to our security” or “our borders should be closed to refugees” ( $M = 4.27$ ,  $SD = 1.70$ ,  $\alpha = .94$ )

**Prior attitudes towards climate change** were measured according to Gustafson (2020), using a semantic-differential scale and four items. For example, we asked whether participants believed that “climate change is NOT happening” vs. “climate change IS happening” ( $M = 4.71$ ,  $SD = 1.46$ ,  $\alpha = .79$ ).

### **Dependent variables**

**Perceived credibility of the visual** was measured with three items, asking participants about the extent to which they found the image in the social media post to be “accurate,” “authentic” and “believable” (Appelman & Sundar, 2016; Oeldorf-Hirsch et al., 2020) ( $M = 3.50$ ,  $SD = 1.68$ ,  $\alpha = .80$ ).

**Agreement with the false claim** portrayed in the image was measured with a single item. Participants were asked to indicate the extent to which they agreed that, to the best of their knowledge, the event depicted in the social media post occurred (see also Wittenberg et al., 2024). Each item *explicitly* referenced the event; for example, “As far as I know, the Friesian Islands are under water due to rising sea levels.” ( $M = 2.85$ ,  $SD = 1.96$ ).

### **Sample**

To determine the necessary sample size, we conducted a simulation-based power analysis in R using the package Superpower (Lakens & Caldwell, 2021). We based our analysis on mean scores reported by Wittenberg et al. (2024), who examined different labels on AI-generated visuals. However, their study did not include added context – as our experiment does – nor did it feature a community note condition. Therefore, our power analysis is only loosely based on existing findings. The results indicated that  $n = 235$  per group, totaling  $N = 940$ , is required to detect the main effect with 90% power. To account for interaction effects with our pre-treatment variables and potential differences across topics covered in the visuals, we planned to oversample, aiming for a total  $N = 1,000$ . Accordingly, we recruited a representative sample of the Dutch

population ( $N=1,018$ ) via the research agency Dynata. (Sample details: Age in years;  $M = 49.49$ ,  $SD = 17.55$ ; Gender: female = 51.6%, male = 48.2%, non-binary = 0.001%; other = 0.001%; Education: low = 24.8%, medium = 42.8%, high = 32.4%; Political orientation, measured on a scale from 0 = “extreme left” to 10 = “extreme right”:  $M = 5.65$ ,  $SD = 2.35$ ; median response time = 7.4 minutes). The survey was fielded between April 7, 2025, and April 28, 2025. To check for randomization of age and political orientation across the four pooled experimental groups we conducted Analysis of Variance (ANOVA), which was successful [Age:  $F(1, 1016) = 0.534$ ,  $p = 0.47$ ; Political orientation:  $F(1, 1016) = 0.128$ ,  $p = 0.72$ ]. We conducted Pearson’s Chi-square tests to check randomization of gender and education, which was also successful [Gender:  $\chi^2(9, N = 1018) = 16.731$ ,  $p = 0.053$ ; Education:  $\chi^2(6, N = 1018) = 0.853$ ,  $p = 0.99$ ].

### ***Manipulation checks***

To assess topic comprehension and distract from key dependent measures, participants rated how much the stimulus addressed “climate” and “migration” (1 = strongly disagree; 7 = strongly agree). As expected, climate change images were significantly more related to climate ( $M = 6.12$ ) than migration images ( $M = 1.65$ ),  $t(1006.40) = 50.69$ ,  $p < .001$ . Conversely, participants rated migration images as significantly more related to migration ( $M = 6.18$ ) than climate change images ( $M = 1.58$ ),  $t(993.94) = 50.56$ ,  $p < .001$ . At the end of the survey, we assessed source label recall. OLS regressions (dummy-coded: 1 = exposed, 0 = others) showed participants were significantly more likely to recall the label source when they had been exposed to it: “Made with AI”  $F(1, 1016) = 172.60$ ,  $p < .001$ ,  $R^2 = .15$ , “Flagged by independent fact-checkers”  $F(1, 1016) = 320.40$ ,  $p < .001$ ,  $R^2 = .24$ , “Readers added context”  $F(1, 1016) = 49.54$ ,  $p < .001$ ,  $R^2 = .05$ , or “no label”  $F(1, 1016) = 117.40$ ,  $p < .001$ ,  $R^2 = .10$ . One-way ANOVAs confirmed that recall was significantly higher in the corresponding exposure condition across all labels (full results in supplementary material B).

### **Results**

To analyze our data, we first used the entire sample ( $N = 1,018$ ), thereby pooling responses across all images. As such, the four condition groups and corresponding interventions encompassed a variety of falsehoods across the topics of climate change and immigration, increasing the generalizability of our findings. To assess whether our results differed by topic, we

then split the sample and repeated the same analysis procedure for the subset of participants who were exposed to visual disinformation regarding climate change ( $n = 502$ ) and immigration ( $n = 516$ ).

### ***Main effects of label source across topics***

To test our first hypotheses **H1ab** – whether the presence of a labeling intervention will result in lower credibility of the visual compared to no intervention, we conducted a one-way analysis of variance (ANOVA). In line with our explanations, the mean score for perceived credibility of the visual was highest in the control condition ( $M = 3.72$ ,  $SD = 1.63$ ) in comparison to an intervention through AI-source ( $M = 3.50$ ,  $SD = 1.68$ ), fact-check source ( $M = 3.46$ ,  $SD = 1.70$ ) and community note ( $M = 3.33$ ,  $SD = 1.69$ ). However, this difference was not statistically significant  $F(3, 1014) = 2.37$ ,  $p = .069$ ,  $\eta^2 = .007$ . The same was the case for our outcome variable agreement with the false claim  $F(3, 1014) = 1.82$ ,  $p = .14$ ,  $\eta^2 = .005$ , where means and standard deviations for each condition were: control group ( $M = 3.07$ ,  $SD = 2.08$ ), AI source ( $M = 2.84$ ,  $SD = 1.92$ ), fact-check source ( $M = 2.76$ ,  $SD = 1.93$ ), and community note source ( $M = 2.69$ ,  $SD = 1.90$ ). Accordingly, we did not conduct post-hoc tests with pairwise mean score comparisons and **reject hypotheses H1ab**. Moreover, we **reject H2ab**, which stated that a fact-checking source will be more effective in lowering credibility of the visual and lowering agreement with the false claim compared to an AI-source. Answering **RQ1** and **RQ2**, we conclude that there is no difference in perceived credibility of the visual and agreement with the labeled false claim between an intervention through an AI source vs. a community note source or between a fact-checking source vs. a community note source. Overall, these results suggest that labels are ineffective when pooling both investigated topics (climate change and immigration).

### ***Interaction effects with pre-existing attitudes towards the source***

To test hypothesis **H4ab**, **H5ab** and **H6ab**, we conducted OLS-based linear regressions using dummy variables with the control condition (no intervention) serving as the reference group. First, we hypothesized that an AI source would be more effective at decreasing credibility of the visual and agreement with the false claim for respondents with higher levels of belief in AI (**H4ab**). This was not the case, as we did not find a significant interaction effect. Specifically, the interaction

between belief in AI and the AI source condition was not significant for either credibility ( $\beta = 0.09, p = .419$ ) or false claim agreement ( $\beta = 0.03, p = .816$ ). Therefore, we **reject H4ab**.

Second, we hypothesized that a fact-checking source is more effective for respondents with higher levels of trust in fact-checking. Here, too, we **reject H5ab**, as the interaction between trust in fact-checking and exposure to the fact-check source was insignificant for perceived credibility ( $\beta = 0.15, p = .236$ ) and false claim agreement ( $\beta = 0.07, p = .659$ ).

Lastly, **H6ab** predicted that a community note source would be more effective in reducing perceived credibility and agreement with false claims among respondents with higher levels of the bandwagon heuristic, i.e., belief in wisdom of the crowds. The interaction between community note source and bandwagon was not significant for perceived credibility, ( $\beta = 0.07, p = .50$ ), nor for agreement with the false claim, ( $\beta = 0.15, p = .236$ ). However, significant positive interactions were found for AI source and bandwagon heuristic ( $\beta = 0.21, p = .05$ ) and fact-checking source and bandwagon heuristic ( $\beta = 0.20, p = .051$ ) for the outcome credibility of the visual, indicating these sources were less effective at higher levels of the bandwagon heuristic. These results provide **partial support for H6a**: the community note *doesn't improve* its effectiveness with higher belief in the wisdom of the crowds, but it *doesn't weaken* it like in the other conditions.

### ***Effects per topic***

To address our research question **RQ3** – to what extent do the underlying thematic categories of AI-generated visual disinformation affect perceived credibility of the visual and agreement with the false claim – we applied the same analytic procedures to the respective subsamples. However, to keep the manuscript concise, we do not report the full set of analyses for hypotheses H1-H6 and research questions RQ1 and RQ2 here, but instead focus on overarching patterns and highlight any divergent findings. We report complete results in our supplementary material C. Moreover, we conducted OLS-based linear regressions to answer **RQ4**, asking to what extent pre-existing attitudes toward (a) climate change or (b) immigration function as a moderator.

***Effects for visual climate change disinformation.*** In short, we find similar patterns for AI-generated visual disinformation for the topic climate change ( $n = 502$ ) as for the analysis with the complete sample. The results suggest that labels are ineffective for AI-generated visual disinformation about climate change, and that this is not moderated by belief in AI, trust in fact-checking or the bandwagon heuristic. To answer **RQ4**, we also did not find any significant

interaction effects with prior attitudes towards climate change for either credibility of the visual  $F(7, 494) = 3.15, p = .003, R^2 = .04$  or belief in the false claim  $F(7, 494) = 2.20, p = .033, R^2 = .03$  ( $ps > .05$ ). A significant main effect of climate change attitudes emerged,  $b = 0.30, SE = 0.10, t = 3.00, p = .003$ , indicating that higher climate concern was associated with greater perceived credibility of the visual across conditions. However, the effectiveness of different labels did not vary based on individuals' attitudes towards the topic of climate change.

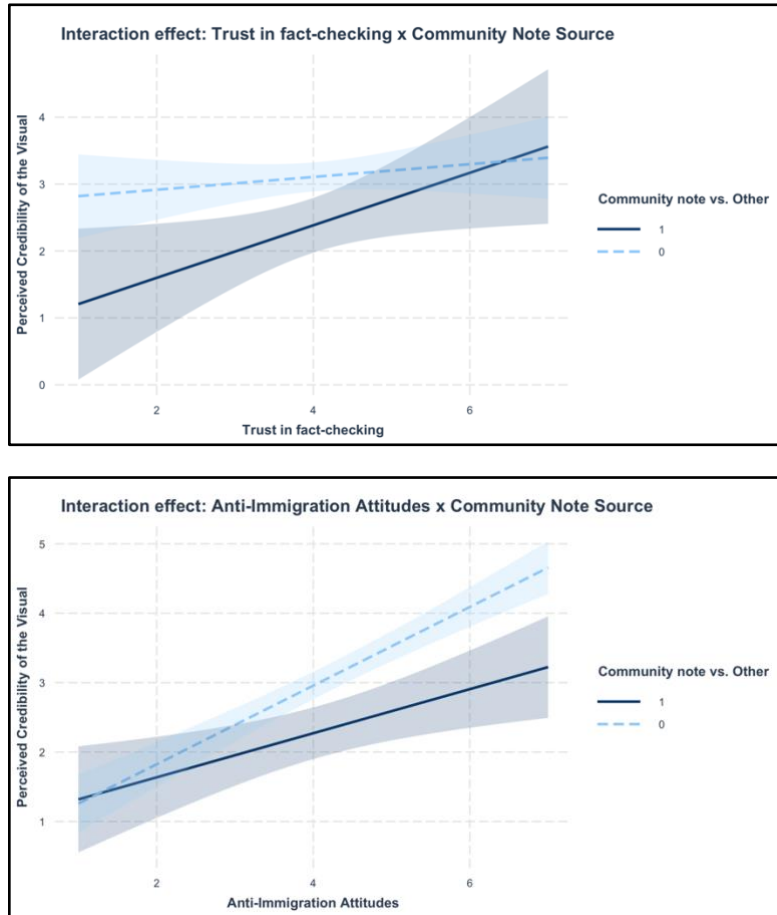
**Effects for visual immigration disinformation.** We repeated the analytical steps for our sub-sample of participants exposed to disinformation about refugees ( $n = 516$ ). Overall, this analysis paints a different picture than when pooling both topics or when considering AI-generated visual climate change disinformation in isolation.

A one-way ANOVA revealed a significant main effect for the outcome perceived credibility of the visual,  $F(3, 512) = 2.89, p = .035$ , partial  $\eta^2 = .017$ . Pairwise t-tests with Bonferroni correction showed no significant differences between conditions (all adjusted  $p > .05$ ), though the difference between the control group ( $M = 3.75, SD = 1.64$ ) and the community note source ( $M = 3.21, SD = 1.70$ ) approached significance ( $p = .057$ ). In addition, a one-way ANOVA revealed a significant effect of source condition on belief in false claims,  $F(3, 512) = 3.71, p = .012$ , partial  $\eta^2 = .022$ . Pairwise comparisons indicated a significant difference between the control ( $M = 3.31, SD = 2.17$ ) and community note conditions ( $M = 2.56, SD = 1.92$ ) ( $p = .02$ ), with a small to medium effect size (Cohen's  $d = 0.37$ ). No other pairwise differences were significant. Overall, these findings suggest that the **community note condition is most effective in reducing belief in false claims about immigration** compared to the control group, and provides tentative evidence that it also reduces the perceived credibility of the visual more.

We also tested various interaction effects for the sub-sample, first considering pre-existing attitudes towards the label source. Similar patterns emerged as in the complete sample when testing for interaction effects with the belief in AI heuristic, which were not significant. Moreover, in the immigration sub-sample, we found a similar significant interaction effect between the AI source and bandwagon heuristic ( $\beta = 0.21, p = .05$ ) and the fact-checking source and bandwagon heuristic ( $\beta = 0.20, p = .051$ ) for the outcome credibility of the visual. Again, this indicates that the community note does not gain effectiveness for those who tend to believe in the wisdom of the crowd, but it doesn't weaken it like in the other conditions. Checking for interaction effects with participants' baseline trust in fact-checking revealed another interesting interaction effect between

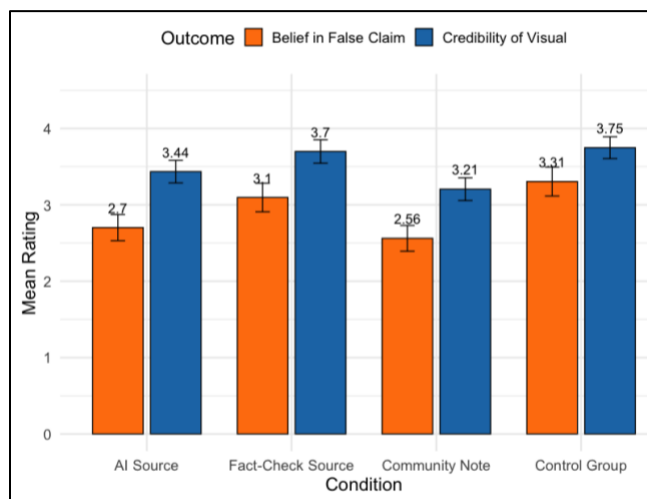
the community note condition and fact-checking trust,  $b = 0.44$ ,  $SE = 0.18$ ,  $t(508) = 2.43$ ,  $p = .015$ . Specifically, greater trust in fact-checking was associated with higher credibility ratings for visuals labeled with community notes. This suggests that **for individuals who have low trust in fact-checks, the community note was particularly successful in lowering the credibility of the false visual portraying the negative consequences of immigration** (see Figure 2). However, no other interaction terms were statistically significant (all  $ps > .07$ ), specifically the one between trust in fact-checking and exposure to a fact-check source.

Answering **RQ4**, we found a significant interaction between the community note condition and anti-immigration attitudes,  $b = -0.29$ ,  $SE = 0.11$ ,  $t(508) = -2.57$ ,  $p = .011$ . Lowering the credibility of the false visual via community notes was stronger among participants with higher anti-immigration attitudes. The main effect of anti-immigration attitudes was also significant,  $b = 0.55$ ,  $SE = 0.08$ ,  $t = 6.97$ ,  $p < .001$ , suggesting that higher anti-immigration attitudes were generally associated with higher perceived credibility of the visuals. Next, we conducted a regression with belief in the false claim as the outcome. The interaction between the community note condition and anti-immigration attitudes approached significance,  $b = -0.25$ ,  $SE = 0.13$ ,  $t(508) = -1.86$ ,  $p = .064$ , indicating a similar trend where community notes might be more effective in reducing belief in false claims among individuals high in anti-immigration attitudes. The main effect of anti-immigration attitudes was again significant,  $b = 0.60$ ,  $SE = 0.10$ ,  $t = 6.28$ ,  $p < .001$ , with higher anti-immigration attitudes predicting stronger belief in the false claim. Overall, this suggests that **community notes are particularly effective for individuals who hold strong anti-immigration attitudes**, if the topic covered in AI-generated visual disinformation is immigration (see Figure 2).

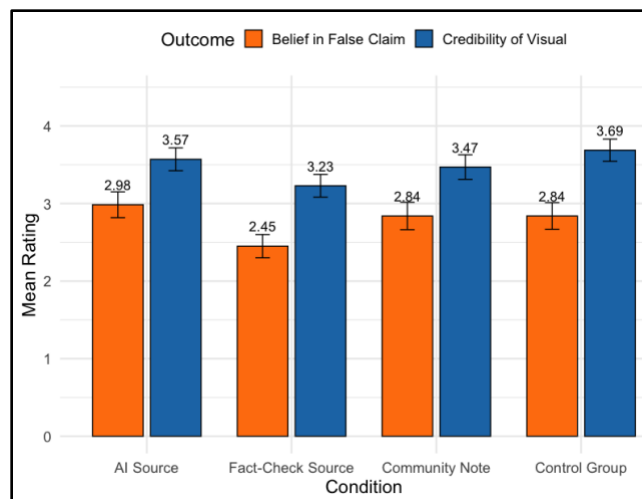


**Figure 2.** Interaction effects of the community note condition with (a) trust in fact-checking and (b) anti-immigration attitudes in the immigration disinformation sub-sample (n = 516), incl. 95% Confidence Intervals

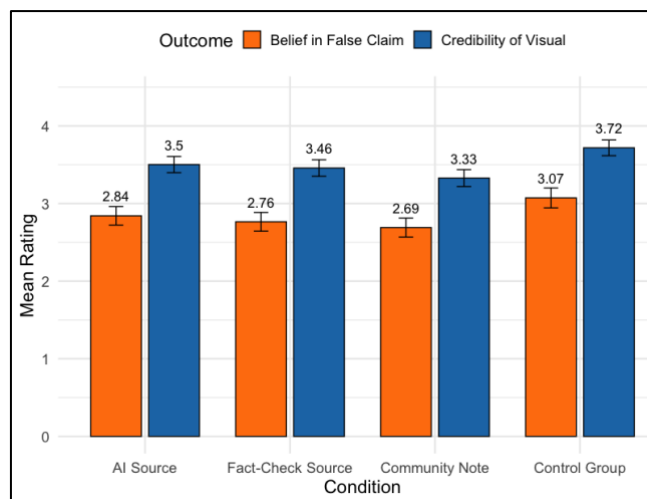
Overall, these findings allow us to answer research question **RQ3**, which asked to what extent the topics of AI-generated visual disinformation play a role for the effectiveness of different platform interventions. The short answer is that topics do matter. When the goal is to correct AI-generated visual disinformation about climate change, none of our interventions turned out to be more successful than the control group without intervention. However, when the AI-generated visual disinformation portrays false narratives around refugees, i.e., covering the topic immigration, the community note is most successful in doing so. It is even more effective for those who already hold negative attitudes towards migrations and those who have low trust in fact-checkers.



(a) Immigration sub-sample



(b) Climate change sub-sample



(c) Complete sample

**Figure 3.** The perceived credibility of AI-generated visual disinformation and belief in the false claim it portrays across conditions and (sub-)samples

***Exploratory analysis (not pre-registered)***

Although the conducted analyses of variance (ANOVAs) did not yield statistically significant main effects based on p-values, we observed meaningful mean differences between the experimental groups as indicated by the effect sizes ( $\eta^2$ -values). Therefore, we proceeded with exploratory, less conservative analyses to further investigate these patterns. Specifically, we employed OLS regression models with k-1 dummy coding, using the control group as the reference

category. This approach does not test between-group differences directly, but rather assesses whether exposure to each specific label, compared to all others combined, predicts the outcome variables. The results revealed that, for the full sample, exposure to the community note significantly reduced both the perceived credibility of the visual and belief in the false claim. This provides tentative evidence that the community note was the most effective intervention overall. **Focusing on the climate change topic specifically, exposure to the fact-check label significantly reduced the credibility of the false visual.** This suggests that while the fact-check label may not have a broad effect, it can be effective within specific contexts. For the immigration topic, the regression results aligned with the ANOVA findings: the community note significantly reduced both outcome measures. Additionally, exposure to the AI-generated label significantly lowered belief in the false claim, indicating that this intervention may also have a positive impact for the right-wing topic. These findings are detailed in Table 1 below.

**Table 1.** OLS-based regressions with dummy variables for the effect of each condition on both outcome variables per sample

Condition	Complete sample		Climate change sub-sample		Immigration sub-sample	
	<i>Dependent variable</i>					
	<i>Credibility of the visual</i> (1)	<i>Belief in false claim</i> (2)	<i>Credibility of the visual</i> (3)	<i>Belief in false claim</i> (4)	<i>Credibility of the visual</i> (5)	<i>Belief in false claim</i> (6)
AI Source	-0.216 (0.147)	-0.231 (0.172)	-0.117 (0.206)	0.145 (0.231)	-0.313 (0.208)	-0.603* (0.252)
Fact-Check Source	-0.260 (0.147)	-0.308 (0.172)	-0.458* (0.205)	-0.389 (0.230)	-0.049 (0.211)	-0.209 (0.255)
Community Note	-0.390** (0.149)	-0.382* (0.175)	-0.217 (0.214)	-0.0004 (0.239)	-0.543** (0.209)	-0.744** (0.252)
Constant (Control Group as reference)	3.718*** (0.104)	3.073*** (0.121)	3.687*** (0.145)	2.840*** (0.163)	3.748*** (0.147)	3.305*** (0.178)
Observations	1,018	1,018	502	502	516	516
R <sup>2</sup>	0.007	0.005	0.011	0.012	0.017	0.021
Adjusted R <sup>2</sup>	0.004	0.002	0.005	0.006	0.011	0.016
Residual Std. Error	1.676 (df = 1014)	1.961 (df = 1014)	1.661 (df = 498)	1.860 (df = 498)	1.686 (df = 512)	2.037 (df = 512)
F Statistic	2.374 (df = 3; 1014)	1.822 (df = 3; 1014)	1.797 (df = 3; 498)	1.968 (df = 3; 498)	2.887* (df = 3; 512)	3.714* (df = 3; 512)

Note: \*p<0.05; \*\*p<0.01; \*\*\*p<0.001

## Discussion

Concerns about the effects of visual disinformation created with the help of generative AI are growing, as cases of AI-generated disinformation, such as deepfakes and image-based disinformation, become more common on social media. However, we know only little about how their potentially detrimental impact could be mitigated, specifically, which platform interventions would reduce the perceived credibility of false visuals, as well as reduce belief in the false claim they portray (Dan, 2025; Wittenberg et al., 2024). Especially since large social media platforms such as Meta and X are withdrawing their efforts in fact-checking and ending their collaboration with journalistic platforms and professional fact-checkers, it is crucial to assess how the effectiveness of different strategies of labelling AI-generated visual content depend on the format of the correction. Although community notes are often promoted and platformed as alternative citizen-driven forms of corrective information, it is important to assess how their effectiveness compares to traditional fact-checking in the context of prominent issues associated with disinformation. In this setting, we relied on a between-subjects experiment in the Dutch setting (N = 1,018) to explore the effectiveness of different corrective messages in the form of labels responding to AI-generated visual disinformation.

Our main findings suggest that, across issues, *no corrective format effectively reduces the credibility of or agreement with visual disinformation*. Unlike prior research on textual corrections (e.g., Clayton et al., 2020; Hameleers & van der Meer, 2020; Walter et al., 2020), we find that AI-generated visuals are more resistant to refutation. Our findings are somewhat in line with Wittenberg et al. (2024), who reported only modest belief-reducing effects from such labels for AI-generated visual disinformation using a larger sample. Similarly, our data show a trend toward lower credibility and false claim belief in label conditions – albeit the lack of statistical significance. Taken together, this leads us to conclude that it is overall more difficult to *meaningfully* reverse the effects of visual disinformation, as was also suggested by earlier work (e.g., Oeldorf-Hirsch et al., 2020; Walter et al., 2020). One explanation for this is that interventions for visuals on social media may be overpowered by the picture and thus overlooked (Oeldorf-Hirsch et al., 2020). However, our manipulation checks confirmed participants recalled the labels correctly. Thus, we contend that the *realism heuristic* and indexicality of visuals may be at play here and hinder critical scrutiny (Sundar, 2008). As visuals are often perceived as direct reflections of reality, labels alone may be insufficient to disrupt the default belief in their authenticity. In

addition, the visual disinformation in our experiment re-iterated the falsehood in both text and image, potentially increasing its content fluency (Prike & Ecker, 2023). A single label may thus not be enough to disrupt the information processing of AI-generated visual disinformation, even if it is registered (see Ecker et al., 2022; Pennycook et al., 2019). Overall, our findings underscore the unique and potentially stronger cognitive impact of visual disinformation which is multifaceted in form, calling for dedicated research attention and platform policies (Weikmann & Lecheler, 2023b).

On a more hopeful note, our results suggest that *platform interventions can be effective, depending on the type of label and topic*. Specifically, in the context of immigration-related visual disinformation, community notes, unlike other corrective formats, reduced belief in false claims and, to some extent, lowered credibility. This effect was strongest among individuals with pronounced anti-immigration attitudes, likely because higher initial agreement with disinformation leaves more room for correction. This aligns with Hameleers and van der Meer (2020), who found that belief congruence can enhance the impact of corrections. Conversely, individuals with weaker anti-immigration views may be less susceptible to misperceptions to begin with, limiting corrective potential. These findings highlight the promise of community notes in addressing polarized issues, as they may circumvent resistance often triggered by institutional fact-checks (e.g., Primig, 2024; Thorson, 2015) by using peer-based cues that are less likely to provoke defensive reactions.

Moreover, the AI-label – often considered a neutral source (Banas et al., 2022) – showed a significant effect in our exploratory regression analysis in reducing beliefs in false claims about immigration. This is in line with findings by Jia et al. (2022), who used a similar set-up to test the effects of textual COVID-19 disinformation. However, in their US-based study, community notes turned out to be particularly ineffective for individuals on the right of the political spectrum. This underlines the need to test the effectiveness of labelling interventions in different contexts. In addition, our exploratory analysis indicates that traditional fact-checks may be more effective for climate-related disinformation, perhaps because individuals more susceptible to it respond better to an established journalistic source (Primig, 2024). In sum, by testing identical interventions across different topics, we show that there is no universal fix for AI-generated visual disinformation, and that platforms must adopt varied, context-sensitive strategies. In our case, bottom-up approaches like community notes and AI-source labels were more effective for right-

leaning topics such as immigration, while top-down interventions from established fact-checking sources may be better suited for left-leaning topics like climate change.

Several limitations should be addressed to further contextualize our findings. First, we fielded our survey experiment in April 2025, not long after Meta’s CEO Mark Zuckerberg announced that he would stop fact-checking efforts on Meta and switch completely to a community note model. This may have paradoxically made right-leaning individuals respond more skeptical of traditional fact-checking efforts, while viewing community notes as more neutral and trustworthy. In line with this, we show that community notes work better for individuals with low trust in fact-checking. However, we did not explicitly mention the word ‘Community Note’ in our design (the label read “*Readers added context*”), so this is only speculative. In addition, while community notes emerged as relatively effective overall, it is important to note that they often reference formal fact-checks in practice (Drolsbach et al., 2024) – an element not accounted for in our design. This raises further questions about the underlying mechanisms of their effectiveness and whether their impact is driven by perceived neutrality, source framing, or specific wording. For instance, the phrase “*Readers added context*” may signal less persuasive intent, making the message more palatable. Thus, a follow-up study is needed to disentangle the effects of the fact-checking content itself from the bottom-up, community-driven format of the notes. In addition, it may be that public sentiment towards the different entities offering content moderation may shift over time. For instance, AI-labels could become more commonplace over time, thus increasing in effectiveness, which warrants future research on this topic (Morosoli et al., 2025).

Moreover, the overall weak effects observed may be partly due to the pooling of multiple images and relatively subtle manipulations. First, the added variation in images led to higher overall standard deviations, increasing prediction error and reducing the likelihood of detecting an effect. In a sense, this reflects a trade-off between measurement precision and external validity, and thus generalizability. Second, in some cases, there may have been limited concrete content to correct, as the visual disinformation used contained some elements of truth (e.g., forest fires have occurred in the Netherlands, and some riots did involve migrants). This partial accuracy, which is another reflection of the external validity of our design (see Hameleers, 2024), may have reduced the potential impact of the interventions. In line with this, our corrections may have only been effective for those individuals prone to be more susceptible in the first place (see Hameleers & van der Meer, 2020). However, it is important to emphasize that all claims were demonstrably

inaccurate, thus contradicting the best available public evidence, and the images *were* manipulated. This strengthens the case that, at least regarding the credibility of the visual, corrective interventions should have had a stronger effect – making the limited impact we find especially noteworthy and allowing for broader generalization than previous single-image studies.

Despite these limitations, our study offers an important step toward understanding how to mitigate the impact of AI-generated visual disinformation on social media. As such disinformation becomes more diverse in form and content, democracies require robust and adaptable interventions to safeguard public discourse. Our findings indicate that labeling interventions can be effective, but that there is no easy fix. Especially at a time when platforms are scaling back content moderation efforts, including fact-checking, we recommend a multifaceted approach, as the effectiveness of interventions may vary depending on the topic and perhaps even the specific image and claim in question.

## **Notes**

1. Link to the pre-registration: <https://osf.io/r38ae>
2. (Minimal) deviations from the pre-registration are reported in supplementary material D.

## **Funding**

This project has received co-funding from the European Union under Grant Agreement number 101158277-BENEDMO-DIGITAL-2023-DEPLOY-04.

## References

- Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, 7(36). <https://doi.org/10.1126/sciadv.abf4393>
- Altay, S., & Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. *PNAS Nexus*, 3(10). <https://doi.org/10.1093/pnasnexus/pgae403>
- Appelman, A., & Sundar, S. S. (2016). Measuring Message Credibility: Construction and Validation of an Exclusive Scale. *Journalism & Mass Communication Quarterly*, 93(1), 59–79. <https://doi.org/10.1177/1077699015606057>
- Banas, J. A., Palomares, N. A., Richards, A. S., Keating, D. M., Joyce, N., & Rains, S. A. (2022). When Machine and Bandwagon Heuristics Compete: Understanding Users' Response to Conflicting AI and Crowdsourced Fact-Checking. *Human Communication Research*, 48(3), 430–461. <https://doi.org/10.1093/hcr/hqac010>
- 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>
- Chadwick, A., & Stanyer, J. (2022). Deception as a Bridging Concept in the Study of Disinformation, Misinformation, and Misperceptions: Toward a Holistic Framework. *Communication Theory*, 32(1), 1–24. <https://doi.org/10.1093/ct/qtab019>
- Chen, Z., Ye, J., Ferrara, E., & Luceri, L. (2025). *Prevalence, Sharing Patterns, and Spreaders of Multimodal AI-Generated Content on X during the 2024 U.S. Presidential Election*. arXiv:2502.11248
- Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A. T., Wolff, A. G., Zhou, A., & Nyhan, B. (2020). Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media. *Political Behavior*, 42(4), 1073–1095. <https://doi.org/10.1007/s11109-019-09533-0>
- Corsi, G., Marino, B., & Wong, W. (2024). The spread of synthetic media on X. *Harvard Kennedy School Misinformation Review*. <https://doi.org/10.37016/mr-2020-140>

- Dan, V. (2025). Deepfakes as a Democratic Threat: Experimental Evidence Shows Noxious Effects That Are Reducible Through Journalistic Fact Checks. *The International Journal of Press/Politics*. <https://doi.org/10.1177/19401612251317766>
- Dan, V., Paris, B., Donovan, J., Hameleers, M., Roozenbeek, J., van der Linden, S., & von Sikorski, C. (2021). Visual Mis- and Disinformation, Social Media, and Democracy. *Journalism & Mass Communication Quarterly*, 98(3). <https://doi.org/10.1177/10776990211035395>
- De Blasio, E., & Selva, D. (2021). Who Is Responsible for Disinformation? European Approaches to Social Platforms' Accountability in the Post-Truth Era. *American Behavioral Scientist*, 65(6), 825–846. <https://doi.org/10.1177/0002764221989784>
- Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes? *International Journal of Press/Politics*, 26(1). <https://doi.org/10.1177/1940161220944364>
- Drolsbach, C. P., Solovey, K., & Pröllochs, N. (2024). Community notes increase trust in fact-checking on social media. *PNAS Nexus*, 3(7). <https://doi.org/10.1093/pnasnexus/pgae217>
- 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*.
- Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1), 13–29. <https://doi.org/10.1038/s44159-021-00006-y>
- Epstein, Z., Fang, M. C., Arechar, A. A., & Rand, D. G. (2023). *What label should be applied to content produced by generative AI?* <https://doi.org/10.31234/osf.io/v4mfz>
- European Commission. (2022). *The Strengthened Code of Practice on Disinformation 2022*.
- European Parliament. (2023, June 18). *EU AI Act: first regulation on artificial intelligence*. <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>
- Gustafson, A., Ballew, M. T., Goldberg, M. H., Cutler, M. J., Rosenthal, S. A., & Leiserowitz, A. (2020). Personal Stories Can Shift Climate Change Beliefs and Risk Perceptions: The

- Mediating Role of Emotion. *Communication Reports*, 33(3), 121–135.  
<https://doi.org/10.1080/08934215.2020.1799049>
- Hameleers, M. (2024). The Nature of Visual Disinformation Online: A Qualitative Content Analysis of Alternative and Social Media in the Netherlands. *Political Communication*, 1–19. <https://doi.org/10.1080/10584609.2024.2354389>
- 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., Humprecht, E., Möller, J., & Lühring, J. (2021). Degrees of deception: the effects of different types of COVID-19 misinformation and the effectiveness of corrective information in crisis times. *Information, Communication & Society*, 1–17.  
<https://doi.org/10.1080/1369118X.2021.2021270>
- Hameleers, M., van der Meer, T. G. L. A., & Dobber, T. (2022). You Won't Believe What They Just Said! The Effects of Political Deepfakes Embedded as Vox Populi on Social Media. *Social Media + Society*, 8(3), 205630512211163.  
<https://doi.org/10.1177/20563051221116346>
- Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public opinion quarterly*, 15(4), 635-650. <https://doi.org/10.1086/266350>
- Instagram Help Center. (2024). *Label AI content on Instagram*. Instagram.  
<https://help.instagram.com/761121959519495>
- Jia, C., Boltz, A., Zhang, A., Chen, A., & Lee, M. K. (2022). Understanding Effects of Algorithmic vs. Community Label on Perceived Accuracy of Hyper-partisan Misinformation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 1–27. <https://doi.org/10.1145/3555096>
- Koch, T. K., Frischlich, L., & Lermer, E. (2023). Effects of fact-checking warning labels and social endorsement cues on climate change fake news credibility and engagement on social media. *Journal of Applied Social Psychology*, 53(6), 495–507.  
<https://doi.org/10.1111/jasp.12959>
- Kreps, S. E., & Kriner, D. L. (2022). The COVID-19 Infodemic and the Efficacy of Interventions Intended to Reduce Misinformation. *Public Opinion Quarterly*, 86(1), 162–175. <https://doi.org/10.1093/poq/nfab075>

- Lakens, D., & Caldwell, A. R. (2021). Simulation-Based Power Analysis for Factorial Analysis of Variance Designs. *Advances in Methods and Practices in Psychological Science*, 4(1), 251524592095150. <https://doi.org/10.1177/2515245920951503>
- Lee, E.-J. (2024). Minding the source: toward an integrative theory of human–machine communication. *Human Communication Research*, 50(2), 184–193. <https://doi.org/10.1093/hcr/hqad034>
- Lee, E.-J., Kim, H. S., Suh, Y. J., & Park, J. W. (2022). Something’s Fishy About It: How Opinion Congeniality and Explainability Affect Motivated Attribution to Artificial Intelligence Versus Human Comment Moderators. *Cyberpsychology, Behavior, and Social Networking*, 25(8), 496–503. <https://doi.org/10.1089/cyber.2021.0347>
- Lee, J., & Hameleers, M. (2024). Effects of Health-related Deepfakes on Misperceptions: Moderating Effects of Issue Relevance and Accuracy Motivation. *Media Psychology*, 1–30. <https://doi.org/10.1080/15213269.2024.2401539>
- Lee, J., Hameleers, M., & Shin, S. Y. (2023). The emotional effects of multimodal disinformation: How multimodality, issue relevance, and anxiety affect misperceptions about the flu vaccine. *New Media & Society*, 146144482311539. <https://doi.org/10.1177/14614448231153959>
- Martel, C., & Rand, D. G. (2023). Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology*, 54, 101710. <https://doi.org/10.1016/j.copsyc.2023.101710>
- Martel, C., & Rand, D. G. (2024). Fact-checker warning labels are effective even for those who distrust fact-checkers. *Nature Human Behaviour*, 8(10), 1957–1967. <https://doi.org/10.1038/s41562-024-01973-x>
- Matich, P., Thomson, T. J., & Thomas, R. J. (2025). Old Threats, New Name? Generative AI and Visual Journalism. *Journalism Practice*, 1–20. <https://doi.org/10.1080/17512786.2025.2451677>
- Meta. (2019, December 16). *Combatting Misinformation on Instagram*. Meta. <https://about.fb.com/news/2019/12/combating-misinformation-on-instagram/>
- Metzger, M. J., Flanagin, A. J., Eyal, K., Lemus, D. R., & Mccann, R. M. (2003). Credibility for the 21st Century: Integrating Perspectives on Source, Message, and Media Credibility in the Contemporary Media Environment. *Annals of the International*

- Communication Association*, 27(1), 293–335.  
<https://doi.org/10.1080/23808985.2003.11679029>
- Morosoli, S., Resendez, V., Naudts, L., Helberger, N., & de Vreese, C. (2025). “I Resist”. A Study of Individual Attitudes Towards Generative AI in Journalism and Acts of Resistance, Risk Perceptions, Trust and Credibility. *Digital Journalism*, 1–20.  
<https://doi.org/10.1080/21670811.2024.2435579>
- Oeldorf-Hirsch, A., Schmierbach, M., Appelman, A., & Boyle, M. P. (2020). The Ineffectiveness of Fact-Checking Labels on News Memes and Articles. *Mass Communication and Society*, 23(5), 682–704. <https://doi.org/10.1080/15205436.2020.1733613>
- Peng, Q., Lu, Y., Peng, Y., Qian, S., Liu, X., & Shen, C. (2024). *Crafting Synthetic Realities: Examining Visual Realism and Misinformation Potential of Photorealistic AI-Generated Images*. <https://doi.org/10.1145/3706599.3719834>
- Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147(12), 1865–1880. <https://doi.org/10.1037/xge0000465>
- Prike, T., & Ecker, U. K. H. (2023). Effective correction of misinformation. *Current Opinion in Psychology*, 54, 101712. <https://doi.org/10.1016/j.copsyc.2023.101712>
- Primig, F. (2024). The Influence of Media Trust and Normative Role Expectations on the Credibility of Fact Checkers. *Journalism Practice*, 18(5), 1137–1157.  
<https://doi.org/10.1080/17512786.2022.2080102>
- Shen, C., Kasra, M., & O’Brien, J. F. (2021). Research note: This photograph has been altered: Testing the effectiveness of image forensic labeling on news image credibility. *Harvard Kennedy School Misinformation Review*. <https://doi.org/10.37016/mr-2020-72>
- Simon, F. M., Altay, S., & Mercier, H. (2023). Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. *Harvard Kennedy School Misinformation Review*. <https://doi.org/10.37016/mr-2020-127>
- Sordo, Z., Chagnon, E., & Ushizima, D. (2025). A Review on Generative AI For Text-To-Image and Image-To-Image Generation and Implications To Scientific Images.  
*arXiv:2502.21151*.
- Sundar, S. S. (2008). *The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility*. 73–100. <https://doi.org/10.1162/dmal.9780262562324.073>

- Taber, C. S., & Lodge, M. (2006). Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science*, 50(3), 755–769. <https://doi.org/10.1111/j.1540-5907.2006.00214.x>
- 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., Hameleers, M., de Vreese, C., Opgenhaffen, M., & Wouters, F. (2024). Beyond Belief Correction: Effects of the Truth Sandwich on Perceptions of Fact-checkers and Verification Intentions. *Journalism Practice*, 1–20. <https://doi.org/10.1080/17512786.2024.2311311>
- Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media and Society*, 6(1). <https://doi.org/10.1177/2056305120903408>
- Vegetti, F., & Mancosu, M. (2020). The Impact of Political Sophistication and Motivated Reasoning on Misinformation. *Political Communication*, 37(5), 678–695. <https://doi.org/10.1080/10584609.2020.1744778>
- 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>
- Weikmann, T., & Lecheler, S. (2023a). 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., & Lecheler, S. (2023b). Visual disinformation in a digital age: A literature synthesis and research agenda. *New Media & Society*, 25(12), 3696–3713. <https://doi.org/10.1177/14614448221141648>
- Weikmann, T., Greber, H., & Nikolaou, A. (2024). After Deception: How Falling for a Deepfake Affects the Way We See, Hear, and Experience Media. *The International Journal of Press/Politics*. <https://doi.org/10.1177/19401612241233539>
- Wittenberg, C., Epstein, Z., Peloquin-Skulski, G., Berinsky, A. J., & Rand, D. G. (2024). *Labeling AI-Generated Media Online*. <https://doi.org/10.31234/osf.io/b238p>
- Yang, Y., Davis, T., & Hindman, M. (2023). Visual misinformation on Facebook. *Journal of Communication*. <https://doi.org/10.1093/joc/jqac051>

## *Supplementary material*

### A. Pre-test:

To make a general claim about the effectiveness of different labelling interventions, we decided to use a variety of AI-generated visual disinformation as stimulus material. In addition, we wanted to make sure that the pool of images did not differ much in itself, even though it consisted of different narratives within and between the topics climate change and immigration. To achieve this, we first created a pool of 30 images displaying various falsehoods in a Dutch context. We did this by asking an AI-image generator (Grok) to create images that support various false claims about the negative consequences of climate change and immigration in the Netherlands. We then pre-tested this initial pool of 30 images with a non-representative sample of Dutch survey participants (N = 100) on the platform Prolific (Sample details: Aged 18 and older; M = 32.54, SD = 8.28; Gender: female = 26%, male = 73%, non-binary = 1%; Education: low = 15%, medium = 48%, high = 37%; Political orientation, measured on a scale from 0 = “extreme left” to 10 = “extreme right”: M = 4.29, SD = 2.14). Each participant evaluated 10 images in succession, rating the following dimensions:

- Plausibility of the claim (7-point scale; 1 item; Wertgen et al., 2021; Wertgen & Richter, 2022),
- Perceived credibility of the visual (7-point scale; 3 items; Appelman & Sundar, 2016; Oeldorf-Hirsch et al., 2020),
- Perceived likelihood that the image was AI-generated (1 item, 10-point scale; 1 = very low, 10 = very high),
- Confidence in their AI-detection judgment (1 item, 7-point scale).

After collecting the data, we calculated the mean scores for each image across all measures and conducted a k-means cluster analysis to identify groups of images with similar perceptual profiles. We tested both a 3-cluster and a 4-cluster solution, which yielded comparable results. Importantly, one cluster consistently included images that were rated as highly plausible and credible and were not suspected to be AI-generated. Another cluster comprised images that were viewed as implausible, low in credibility, and highly suspected to be AI-generated. These

clusters were stable across both clustering solutions. Based on this, we excluded the images from these two clusters from further use. Instead, we selected the remaining images, which were more uniformly evaluated across the measured variables, to be used in our main study.

### B. ANOVA results for manipulation check

#### Analysis of variance for manipulation check question 1 – “**AI-label**”

A one-way ANOVA was conducted to test whether recognition of the AI label differed across experimental conditions. There was a significant effect of condition on AI-label recognition,  $F(3, 1014) = 60.33, p < .001$ , indicating that participants’ perceptions varied significantly depending on the type of labeling intervention. Specifically, participants in the AI-label condition reported significantly higher recognition of the label “Made with AI” ( $M = 4.21, SD = 1.12$ ) compared to all other conditions, according to post-hoc comparisons with Bonferroni-corrections:

- **Community Note:**  $p < .001$
- **Control Group:**  $p < .001$
- **Fact-Check Label:**  $p < .001$

All means, standard deviations, n per group and standard errors can be found in the table below:

<b>Condition</b>	<b><i>M</i></b>	<b><i>SD</i></b>	<b><i>n</i></b>	<b><i>SE</i></b>
AI Label	4.21	1.12	259	0.07
Fact-Check Label	3.10	1.48	255	0.09
Community Note	2.92	1.47	242	0.09
Control Group	2.77	1.31	262	0.08

#### Analysis of variance for manipulation check question 2 – “**Fact-check label**”

A one-way ANOVA was conducted to test whether recognition of the fact-check label differed across experimental conditions. There was a significant effect of condition on AI-label recognition,  $F(3, 1014) = 106.8, p < .001$ , indicating that participants’ perceptions varied significantly depending on the type of labeling intervention. Specifically, participants in the fact-

check label condition reported significantly higher recognition of the label “Flagged by independent fact-checkers” ( $M = 3.97, SD = 1.21$ ) compared to all other conditions, according to post-hoc comparisons with Bonferroni-corrections:

- **Community Note:**  $p < .001$
- **Control Group:**  $p < .001$
- **Fact-Check Label:**  $p < .001$

All means, standard deviations, n per group and standard errors can be found in the table below:

<b>Condition</b>	<b><i>M</i></b>	<b><i>SD</i></b>	<b><i>n</i></b>	<b><i>SE</i></b>
AI Label	2.37	1.24	259	0.08
Fact-Check Label	3.97	1.21	255	0.08
Community Note	2.43	1.24	242	0.08
Control Group	2.42	1.15	262	0.07

#### Analysis of variance for manipulation check question 3 – “**Community note**”

A one-way ANOVA was conducted to test whether recognition of the community note differed across experimental conditions. There was a significant effect of condition on AI-label recognition,  $F(3, 1014) = 19.21, p < .001$ , indicating that participants’ perceptions varied significantly depending on the type of labeling intervention. Specifically, participants in the community note condition reported significantly higher recognition of the label “Users added context” ( $M = 3.14, SD = 1.21$ ) compared to all other conditions, according to post-hoc comparisons with Bonferroni-corrections:

- **Community Note:**  $p < .001$
- **Control Group:**  $p < .001$
- **Fact-Check Label:**  $p < .001$

All means, standard deviations, n per group and standard errors can be found in the table below:

<b>Condition</b>	<b><i>M</i></b>	<b><i>SD</i></b>	<b><i>n</i></b>	<b><i>SE</i></b>
AI Label	2.33	1.18	259	0.07
Fact-Check Label	2.60	1.27	255	0.08
Community Note	3.14	1.32	242	0.08
Control Group	2.58	1.12	262	0.07

Analysis of variance for manipulation check question 3 – “**Control group**”

A one-way ANOVA was conducted to test whether recognition of the control group differed across experimental conditions. There was a significant effect of condition on AI-label recognition,  $F(3, 1014) = 74.79, p < .001$ , indicating that participants’ perceptions varied significantly depending on the type of labeling intervention. Specifically, participants in the control group reported significantly higher recognition that there was no label ( $M = 3.49, SD = 1.21$ ) compared to all other conditions, according to post-hoc comparisons with Bonferroni-corrections:

- **Community Note:**  $p < .001$
- **Control Group:**  $p < .001$
- **Fact-Check Label:**  $p < .001$

All means, standard deviations, n per group and standard errors can be found in the table below:

<b>Condition</b>	<b><i>M</i></b>	<b><i>SD</i></b>	<b><i>n</i></b>	<b><i>SE</i></b>
AI Label	2.37	1.38	259	0.09
Fact-Check Label	2.35	1.38	255	0.09
Community Note	2.61	1.39	242	0.09
Control Group	3.49	1.24	262	0.08

C. Complete hypothesis testing for sub-samples

Sample details **climate change** topic ( $n = 502$ ): Aged 18 and older;  $M = 49.45$ ,  $SD = 17.69$ ; Gender: female = 49.4%, male = 47.5%, non-binary = 0.001%; other = 0.001%; Education: low = 23.3%, medium = 42.6%, high = 31.4%; Political orientation, measured on a scale from 0 = “extreme left” to 10 = “extreme right”:  $M = 5.61$ ,  $SD = 2.29$ ; mean response time = 7.9 minutes;

We conducted a one-way analysis of variance (ANOVA) to test if a labeling intervention will result in lower credibility of the visual compared to no intervention in the climate change sub-sample. The ANOVA did not provide significant results for the outcome variable credibility of the visual  $F(3, 498) = 1.80$ ,  $p = .147$ ,  $\eta^2 = 0.01$  nor for agreement with the false claim  $F(3, 498) = 1.97$ ,  $p = .118$ ,  $\eta^2 = 0.01$ . Accordingly, we did not conduct post-hoc tests with pairwise mean score comparisons and reject hypotheses H1ab for this sub-sample. Moreover, we reject H2ab, which stated that a fact-checking source will be more effective in lowering credibility of the visual and lowering agreement with the false claim compared to an AI-source. Answering RQ1 and RQ2, we conclude that there is no difference in perceived credibility of the visual and agreement with the labeled false claim between an intervention through an AI source vs. a community note source or between a fact-checking source vs. a community note source. Overall, these results suggest that labels are ineffective for AI-generated visual disinformation about climate change. The means and standard deviations per experimental group are reported in the table below.

**Table.** Mean scores, standard deviations and standard errors

<b>Condition</b>	<i>Credibility of the visual</i>			<i>Belief in false claim</i>			<i>n</i>
	<i>M</i>	<i>SD</i>	<i>SE</i>	<i>M</i>	<i>SD</i>	<i>SE</i>	
AI Label	3.75	1.66	0.15	2.98	1.88	0.17	128
Fact-Check Label	3.23	1.67	0.15	2.45	1.72	0.15	131
Community Note	3.47	1.68	0.16	2.84	1.87	0.18	112
Control Group	3.69	1.64	0.14	2.84	1.97	0.15	131

To test whether baseline attitudes towards any of the sources would moderate the effects on credibility of the visual or belief the false claim, we conducted OLS-based linear regressions using dummy variables with the control condition (no intervention) serving as the reference group. We did not find a significant interaction effect between belief in AI and the AI source condition for either credibility ( $\beta = 0.09, p = .419$ ) or false claim agreement ( $\beta = -0.034, p = .84$ ). We did not find a significant interaction effect between trust in fact-checking and exposure to a fact-checking source for credibility of the visual ( $\beta = 0.002, p = .99$ ) or false claim agreement ( $\beta = -0.16, p = .43$ ). Lastly, we did not find a significant interaction effect between the bandwagon heuristic and exposure to a community note for credibility of the visual ( $\beta = 0.08, p = .59$ ) or false claim agreement ( $\beta = -0.15, p = .40$ ). We reject H4ab, H5ab and H6ab for the climate change sub-sample.

Sample details **immigration** topic ( $n = 516$ ): Aged 18 and older;  $M = 49.53, SD = 17.43$ ; Gender: female = 50.8%, male = 48.8%; Education: low = 23.9%, medium = 43.8%, high = 32.3%; Political orientation, measured on a scale from 0 = “extreme left” to 10 = “extreme right”:  $M = 5.68, SD = 2.40$ ; mean response time = 6.8 minutes;

We conducted a one-way analysis of variance (ANOVA) to test if a labeling intervention will result in lower credibility of the visual compared to no intervention in the immigration sub-sample. The ANOVA for the outcome variable credibility of the visual was significant:  $F(3, 512) = 2.89, p = .035, \eta^2 = 0.017$ . Pairwise t-tests with Bonferroni correction showed no significant differences between conditions (all adjusted  $p > .05$ ), though the difference between the control group ( $M = 3.75, SD = 1.64$ ) and the community note source ( $M = 3.21, SD = 1.70$ ) approached significance ( $p = .057$ ). In addition, a one-way ANOVA revealed a significant effect of source condition on belief in false claims,  $F(3, 512) = 3.71, p = .012, \text{partial } \eta^2 = .022$ . Pairwise comparisons indicated a significant difference between the control ( $M = 3.31, SD = 2.17$ ) and community note conditions ( $M = 2.56, SD = 1.92$ ) ( $p = .02$ ), with a small to medium effect size (Cohen’s  $d = 0.37$ ). No other pairwise differences were significant.

**Table.** Mean scores, standard deviations and standard errors

<b>Condition</b>	<b><i>Credibility of the visual</i></b>			<b><i>Belief in false claim</i></b>			<b><i>n</i></b>
	<b><i>M</i></b>	<b><i>SD</i></b>	<b><i>SE</i></b>	<b><i>M</i></b>	<b><i>SD</i></b>	<b><i>SE</i></b>	
AI Label	3.44	1.69	0.15	2.70	1.96	0.17	131
Fact-Check Label	3.70	1.71	0.15	3.10	2.09	0.19	124
Community Note	3.21	1.70	0.15	2.56	1.92	0.17	130
Control Group	3.75	1.64	0.14	3.31	2.17	0.19	131

Checking for interaction effects with participants' baseline trust in fact-checking revealed another interesting interaction effect between the community note condition and fact-checking trust,  $b = 0.44$ ,  $SE = 0.18$ ,  $t(508) = 2.43$ ,  $p = .015$ . Specifically, greater trust in fact-checking was associated with higher credibility ratings for visuals labeled with community notes. This suggests that for individuals who have high trust in fact-checks, the community note was slightly less successful in lowering the credibility of the false visual portraying the negative consequences of immigration (see Figure 2). However, no other interaction terms were statistically significant (all  $ps > .07$ ), specifically the one between trust in fact-checking and exposure to a fact-check source.

#### D. Deviations from the pre-registration

We deviate from our pre-registration as we (1) excluded a pre-registered part from the manuscript and (2) included an exploratory part in the manuscript. For transparency, these deviations are reported here.

1. In the pre-registration, we outlined expectations concerning the baseline perceived source credibility of the different labelling sources. These include:

H: The presence of a labeling intervention through a fact-checking source will be perceived as more credible compared to an intervention through an AI source.

RQ: Is there a difference in perceived credibility of the label source between a fact-checking source vs. a community note source or RQ: between an AI source vs. a community note source?

We decided not to include these in our manuscript, as we did not find that the findings would enhance the contribution of the paper. However, for transparency, we report the hypothesis testing here. As such, we conducted a one-way ANOVA to examine the effect of source condition on source credibility ratings. Specifically, we asked to what extent participants found the source indicating the AI-manipulation to be qualified, trustworthy, competent, knowledgeable or biased (1 = strongly disagree; 7 = strongly agree) (see Dobber et al., 2023). The results of the ANOVA indicated that there was no significant effect of condition on credibility ratings,  $F(2, 753) = 1.43, p = .24, \eta^2 = .004$  based on the following mean differences: AI-label ( $M = 3.66, SD = 1.50$ ), fact-check label ( $M = 3.60, SD = 1.52$ ) and community note source ( $M = 3.44, SD = 1.44$ ). The control group data was excluded from the analysis due to missing values. The findings overall suggest that there is no difference in perceived source credibility between an AI-label, fact-check label or community note.

Dobber, T., Kruikemeier, S., Votta, F., Helberger, N., & Goodman, E. P. (2025). The effect of traffic light veracity labels on perceptions of political advertising source and message credibility on social media. *Journal of Information Technology & Politics*, 22(1), 82–97. <https://doi.org/10.1080/19331681.2023.2224316>

2. Moreover, in our pre-registration, the part about the differences in topic was initially pre-registered as exploratory and read:

RQ: To what extent do the underlying thematic categories of AI-generated visual disinformation (climate change vs. immigration) affect the perceived credibility of the visual, perceived credibility of the label source and agreement with the false claim?

RQ: To what extent is this moderated by pre-existing attitudes towards either (a) climate change or (b) immigration?

We included this part in the main paper (see above).