TOPICAL POLICY BRIEF

Labeling AI-Generated Content: Promises, Perils, and Future Directions

Chloe Wittenberg
Ziv Epstein
Adam J. Berinsky
David G. Rand

November 28, 2023

MIT Schwarzman College of Computing
MIT Washington Office
aipolicy.mit.edu

Authors

Chloe Wittenberg
Postdoctoral Associate
Department of Political Science

Ziv Epstein
Research Assistant
Media Lab

Adam J. Berinsky
Mitsui Professor of Political Science
Department of Political Science

David G. Rand
Professor of Management Science and Brain and Cognitive Sciences
MIT Sloan School of Management
Department of Brain and Cognitive Sciences

For correspondence: drand@mit.edu

Executive Summary

- Labeling is a commonly proposed strategy for reducing the risks of generative artificial intelligence (AI). This approach involves applying visible content warnings to alert users to the presence of AI-generated media online (e.g., on social media, news sites, or search engines).
- Although there is little direct evidence regarding the effectiveness of labeling AI-generated media, a large academic literature suggests that warning labels can substantially reduce belief in, and sharing of, content debunked by professional fact-checkers. Thus, there is reason to believe that labeling could help inform members of the public about AI-generated media.
- In this paper, we provide a framework for helping policymakers, platforms, and practitioners weigh various factors related to the labeling of AI-generated content online.
- First, we argue that, before developing labeling programs and policies related to generative AI, stakeholders must establish the objective(s) that labeling is intended to accomplish. Here, we distinguish two such goals:
    - Communicating to viewers the *process* by which a given piece of content was created or edited (i.e., with or without using generative AI tools).
    - Diminishing the likelihood that content *misleads* or *deceives* its viewers (a result that does not necessarily depend on whether the content was created using AI).
- We then highlight several important issues and challenges that must be considered when designing, evaluating, and implementing labeling policies and programs, including the need to:
    - Determine what types of content to label and how to reliably identify this content at scale.
    - Consider the inferences viewers will draw about both labeled *and* unlabeled content.
    - Evaluate the efficacy of labeling approaches across contexts, including different media formats, countries, and sub-populations.

Introduction

Rapid improvements in the sophistication and accessibility of generative artificial intelligence (AI) have made it easier than ever for creators to produce realistic videos, images, and audio of almost anything imaginable—including events that never took place. Although generative AI has myriad applications, its potential to amplify the production and dissemination of audiovisual misinformation has become a source of widespread concern (*1–3*). Members of the American public by and large report limited experience with generative AI tools (*4*, *5*), and many say that altered videos and images sow confusion about the facts of current events and issues (*6*). Some academic research likewise finds that members of the public often struggle to discern whether AI-generated media are real or fabricated (*7–9*) and these challenges will undoubtedly worsen as technology continues to improve and evolve (*10*, *11*).

One proposed response to this challenge involves visibly labeling AI-generated media (e.g., with text or graphics superimposed on, or displayed next to, the media), to inform users that the content was created using generative AI. This approach has already attracted legislative attention in both the United States and abroad (*12–15*); for example, the bipartisan AI Labeling Act of 2023 (*16*) would require developers of generative AI systems to include "clear and conspicuous" disclosures indicating that content was produced using AI.[1] At a White House summit in July 2023 (*17*), seven leading technology companies also pledged to develop "robust technical mechanisms" (*18*) for communicating to users when content is AI-generated. And early attempts at such labeling are already appearing across social media platforms (*1*); in recent months, platforms including Instagram (*19*), Google (*20*), and TikTok (*21*) have begun to explore strategies for alerting users to content that has been created or modified using AI.

This emphasis on labeling AI-generated content should be considered in light of academic work demonstrating the utility of warning labels in addressing misleading and deceptive content online. In particular, past studies found that the addition of fact-checking labels to posts identified as misinformation can reduce people's likelihood of believing (*22–26*) and sharing (*24*, *27–29*) false and unsubstantiated claims (*30*). Labeling AI-generated media online may likewise be able to temper the negative effects of this content, but additional research is needed to understand both when and how such labels should be applied.

In this paper, we therefore lay out a framework aimed at helping policymakers and platforms navigate key questions associated with labeling AI-generated media online. In the first section, we underscore the role of institutional *goals* and *strategy* in shaping the design and implementation of a labeling policy or program. In the second section, we then outline several *considerations* and *challenges* related to labeling, including (i) identifying the correct subset of content to label, (ii) mitigating potential unintended consequences of labeling for broader media trust, and (iii) ensuring the efficacy and generalizability of labels across contexts. Throughout, we discuss strategies for labeling audiovisual media (including videos, audio, images, and graphics), as other work has examined the distinct, but related, domain of AI-generated text (*2*). Nevertheless, we take a broad view of the field—focusing on a wide variety of settings in which labeling could occur (e.g., on social media, news platforms, or search engines). We thereby highlight ways that labeling may be able to contribute to core policy objectives, while also emphasizing opportunities for additional research and analysis on this emerging topic.

[1] An open question is how best to impose and enforce labeling requirements. As we discuss throughout this paper, labeling can take many different forms. Which governmental agencies or groups are responsible for overseeing labeling efforts may fundamentally depend on the scope of this activity; for instance, generalized systems aimed at flagging all forms of AI-generated content may require different types of agency involvement than more focused attempts to regulate the use of AI in, say, political advertising. At the same time, it is important to ensure consistency in labeling formats across contexts; a proliferation of different label types and standards will only engender confusion.

A Framework for Labeling AI-Generated Media

Despite widespread interest in the use of labeling to mitigate the influence and spread of AI-generated media online, it remains unclear what term(s) should be applied to this content. AI-generated media are far from monolithic; different types of content vary in their manner and degree of algorithmic intervention, as well as their likely consequences for society at large. Moreover, members of the public possess varying levels of familiarity with and knowledge about AI and related technologies (*4*, *31*), raising questions about whether, and under what conditions, specific labels will be comprehensible to a general audience. How policymakers and technology companies resolve these issues should depend greatly on their overarching goals and organizational strategies. In short, the design, evaluation, and implementation of any labeling policy or program must reflect the core objective(s) that this labeling is intended to accomplish.

To this end, we outline two goals that could motivate the labeling of AI-generated media online. The first is a "process-based" goal that focuses exclusively on the technical processes through which a given piece of content was created and/or modified. From this perspective, the primary function of an AI labeling program is to communicate to users how a particular piece of content was produced, while remaining agnostic about its potential consequences for viewers or society more generally. To achieve this goal, a labeling process might seek to identify and flag any content that was made or edited using generative AI technology, regardless of its format (e.g., video, audio, or image), domain (e.g., politics, art, or science), or conceivable impact on viewers' beliefs and behavior.[2]

The second, in contrast, is an "impact-based" goal centered on the content's potential harm. Much of the prevailing discourse surrounding the labeling and detection of AI-generated media is grounded in fears that such content could mislead or otherwise deceive members of the public. For example, recent calls to regulate the use of AI in political advertising (*11*, *33*, *34*) stem from concerns that "deepfakes" could sway voting behavior, alter election outcomes, or incite political violence (*3*, *35*). These concerns about public deception are not unique to politics; indeed, the proliferation of generative AI may have even more widespread consequences in other areas, including the perpetration of fraud and scams (*36*, *37*) and the creation of nonconsensual sexual imagery (*38*). Labeling efforts could therefore focus on the extent to which content is likely to mislead—for instance, by integrating tools for identifying AI-generated media with existing systems aimed at detecting misinformation (which many technology

[2] As noted above, concerns about misinformation typically underlie calls to label AI-generated content. However, there are several reasons why one might favor a process-based approach. The first is *philosophical*: there may be inherent value in establishing and communicating content provenance in a transparent and accessible way. The second is *practical*: while it can be tricky to conclusively discern whether content is true or false, it may be easier (and less contentious) for platforms to reliably estimate the processes by which this content was created. The third is about *credibility*: at a time when attitudes toward fact-checkers are highly polarized (*32*), people may be more open to and accepting of "neutral" process labels, compared to labels that resemble more traditional fact-checks.

companies have already made substantial investments in, e.g., professional fact-checkers, crowd raters, and machine learning algorithms (*39*–*41*)).

Although these two goals are not mutually exclusive, they are clearly distinct. As summarized in Table 1, not all AI-generated media are inherently misleading (e.g., digital art (*42*)), nor are all forms of misleading content produced using artificial intelligence (*43*). For example, so-called "cheapfakes" that present audio or visuals out of context or use conventional editing techniques to misconstrue events may be just as damaging to the information environment as entirely synthetic media constructed using generative AI (*44*, *45*). Moreover, recent work finds that labeling news headlines as AI-generated diminishes viewers' belief in both true *and* false headlines (*46*, *47*). These results suggest that different types of labeling goals may be in tension. In particular, whereas process-based labels, by design, tend to be transparent about media's provenance, they may be less informative about veracity—and therefore have the potential to create confusion about whether presented information is true or false.

Aligning Goals, Design, and Implementation

A labeling program's objectives also directly inform its design and implementation. First, different goals demand fundamentally different framing and conceptualization. Process-based labels that report methods of content creation would likely need to adopt a neutral stance, as a large portion of AI-generated media is not intrinsically malicious or deceptive in nature. Conversely, impact-based labels denoting misleading (or inaccurate/disputed) content might instead carry a distinctly negative connotation, given their end goal of reducing belief in and engagement with the labeled content. Calibrating the tone of labels is crucial to ensuring they are interpreted as intended. Following this logic, different goals may also shape the consequences of labeling for a post's subsequent distribution. On the one hand, process-based labels should likely not affect the algorithmic ranking of content on social media, given that these labels could be applied to harmless or even beneficial posts (*46*). On the other hand, consistent with existing practice (*48*), AI-generated content labeled as false or misleading should perhaps also be downranked, so as to minimize its likelihood of reaching a wide audience.

At a more basic level, the wording used to label content should also be precise about what types of content are—and, just as importantly, are *not*—covered by a specific label. In surveys conducted by several members of our research team (*49*), we asked a large sample of people in the United States, Brazil, India, Mexico, and China to indicate how well nine common labeling terms described twenty types of content that varied both in their production process and their potential to mislead (for examples, see Table 1). Overall, we found that different terms satisfy different aims. On the one hand, members of the public most consistently associated the terms "AI-generated," "Generated with an AI tool," and "AI-manipulated" with content that was constructed using AI technology—regardless of whether this content was misleading. As such, if labeling policies and programs seek to transparently communicate the processes by which AI-generated media were created, terms that explicitly reference AI may be most appropriate. On the other hand, terms like “Manipulated” and “Deepfake” were most consistently associated with misleading content, regardless of the methods through which this content was generated, making these terms better suited to labeling strategies aimed at identifying deceptive media.[3]

Importantly, however, participants in this study reported that both types of labels—that is, both process- *and* impact-based labels—would make them less confident that the events shown in the presented media took place. Thus, if labels are applied to AI-generated content regardless of its misleadingness, this is likely to reduce belief in content that is accurate but AI-generated—a possibility that is supported by other experiments assessing the causal effect of process-based labels on the perceived accuracy of news headlines (*46*, *47*). Together, these results highlight the importance of well-defined objectives in shaping how and where generative AI disclosures are implemented.

Considerations and Challenges Related to Labeling

Even if labeling policies and programs are developed with a clear set of goals in mind—and language to match—several key challenges remain.

Identifying Which Content to Label

First and foremost is the question of *what content to label*. Existing efforts to identify misleading content on social media typically rely on professional fact-checkers (*39*) and/or the “wisdom of crowds” (*50*), along with machine learning classifiers. However, as attention has shifted toward generative AI, questions remain about how best to determine whether content has been created or manipulated using these specific technologies. As platforms start to roll out AI labeling programs, many have relied on creators to voluntarily report their use of AI tools (*1*, *21*). Such self-disclosures are easy to implement and require minimal top-down enforcement by platforms, but they are unlikely to be adopted by the very actors whose use (or abuse) of generative AI tools is expected to be most harmful. Thus, reliance on voluntary self-disclosure will be insufficient for addressing the potential risks associated with AI-generated content.

Computational approaches to detecting AI-generated media (e.g., using machine learning and forensic analysis (*7*)) may be able to circumvent some of these impediments to scalability. These methods identify statistical patterns and artifacts in AI-generated media, allowing for post-hoc detection of media manipulation. Nevertheless, such systems can fail to uncover new forms of AI-generated media designed to avoid detection (*42*) and typically produce probabilistic, rather than definitive, estimates of whether a given piece of content was made using generative AI. These dynamics are potentially problematic, as misclassification of AI-generated media as authentic (or vice

[3] These results also point to a striking discrepancy in people’s interpretations of the terms “AI-manipulated” and “Manipulated.” One possible explanation for this pattern is that the inclusion of references to AI may prompt individuals to primarily focus on whether content was algorithmically generated, without attending to other features of this content.

		Generation Process	
		AI-Generated	**Not AI-Generated**
Potential to Mislead	**Misleading**	● *A video created from scratch using a computer algorithm that depicts events that never actually occurred* ● *An image in which one person's face is replaced with another person's using a computer algorithm*	● *A video that a person edited or spliced together to present existing footage out of context* ● *An image that was paired with an inaccurate title or caption to misrepresent the presented content*
	Not Misleading	● *A piece of digital art created from scratch by feeding text prompts to a computer algorithm* ● *An image that was edited using a computer algorithm to remove strangers from the background of a family photo*	● *A piece of digital art that was created by a person drawing on a touchscreen and using software to color, shade, and add final touches to their work* ● *An image whose brightness and contrast was adjusted using a photo editing application*

Table 1: Examples of online content that vary in both their *production process* (i.e., AI-generated or not) and their *potential risks* (i.e., misleading or deceiving viewers).

versa) could undermine users' confidence in the legitimacy of labeling efforts and erode trust in the media ecosystem more generally (*51*). Hybrid strategies that blend machine learning, forensic techniques, and crowdsourcing (*52*) may help improve the robustness of AI detection systems but are still vulnerable to the emergence of new forms of AI-generated content. These automated approaches may therefore require continual fine-tuning and adaptation to ensure they are able to keep pace with advancements in AI technology.

A final set of strategies, in contrast, focuses on more direct disclosure methods. These techniques embed signals about whether content is authentic or AI-generated into the content itself.[4] For instance, digital signature-based approaches encode information about the origins of a piece of content, or its "provenance," via a cryptographically secure chain-of-custody (e.g., following the Coalition for Content

[4] Many of these mechanisms are commonly described as "digital watermarking" in the popular press, though they span a wide range of potential approaches to providing disclosures (e.g., via metadata, an imperceptible digital hash, or a front-end label that is visible to users). Although distinguishing these different systems is beyond the scope of this short paper, this subject has been widely discussed by both experts and practitioners (*53*, *54*).

Provenance and Authenticity, or C2PA, protocol). Because these digitally signed statements can be generated at the point of creation, they are less prone to manipulation and evasion and can overcome previously mentioned obstacles related to misclassification and self-disclosure (*42*). While this approach comes with coordination challenges related to the development of industry-wide standards and the need for widespread adoption across companies, governments are well-positioned to address these challenges through regulatory mandates.

Indirect Effects of Labeling

Second, the effects of labels may extend far *beyond the individual pieces of content* to which they are applied. Most notably, wide-ranging efforts to draw attention to AI-generated content could predispose the public to question the veracity of authentic content (*55*, *56*). Recent research on AI disclosures finds that people tend to be less trusting of content tagged as AI-generated, regardless of its underlying veracity or provenance (*46*, *47*). In this same vein, some research suggests that general warnings (as opposed to labels attached to specific pieces of content, e.g. interventions that educate users about "deepfakes" (*57*)) can make people skeptical of *all* media they encounter, leading them to erroneously discount real information (*22*, *58*). In addition, past research has uncovered an "implied truth effect" (*24*), where the application of fact-checking warnings increases the perceived credibility of unlabeled content, even in instances when this content is, in fact, untrustworthy. It is possible that a similar "implied authenticity effect" might occur when labeling AI-generated media, particularly in the absence of an analogous system for identifying and validating content created without AI. When assessing the impacts of a labeling program, it is critical to bear in mind not just how labels influence individuals' responses to tagged content but also how they affect inferences about unlabeled posts and about the media environment more generally.

Labeling may also have consequences for users' beliefs and expectations. In particular, conspicuous efforts to apply labels may convey to users that AI-generated media (and/or misinformation) are widespread, thereby inadvertently normalizing the dissemination of this content. In addition, the introduction of a novel warning system may initially capture viewers' attention and interest, but users may gradually become inured to labels over time, thereby diminishing their long-term potency (a form of "banner blindness," see (*59*, *60*)). Additional research is needed to better understand whether the immediate, direct benefits of labeling in mitigating harm outweigh the long-term, indirect effects of labeling on broader attitudes and behavior.

Contextual Differences

Finally, different *contexts* may necessitate different labeling approaches. Although we focus here on the domain of audiovisual media (including video, audio, images, and graphics), AI-generated text is an important—and closely related—phenomenon with its own unique features and challenges (*2*, *61*). Further, not all users may interpret labels in the same way. As just one example, in our cross-national study examining the comprehensibility of various labeling phrases (*49*), participants from China interpreted the term “artificial” very differently than participants in other countries, reflecting linguistic differences in the types of behavior this word connotes outside the domain of artificial intelligence.[5] Implementing a labeling program at scale requires close attention to these cultural and semantic distinctions, especially in light of the global reach of generative AI and the international user base of many online platforms.

Conclusion

At a time when generative AI systems are increasingly capable of fabricating high-quality media, the visible and transparent labeling of AI-generated content offers one potential safeguard against deception and confusion. As policymakers, technology companies, academics, and other actors debate strategies for AI disclosures, it is vital that they be clear about the objectives of such disclosures—which may include a desire to convey the processes through which content was created, a desire to identify misleading content, or some combination of these and/or other goals. These objectives can provide a foundation for determining (i) what types of content to label and (ii) how to design labels that are both accurate and credible to a wide audience. When establishing policy guidelines and programmatic strategies, stakeholders should also remain attuned to the consequences of disclosures not just for tagged content but also untagged content, given the risk that a fragmented or unreliable labeling system could engender mistrust and further blur the lines between reality and fiction. As artificial intelligence systems continue to evolve at a whirlwind pace, it is imperative for policymakers and platforms to carefully weigh these considerations when regulating, designing, evaluating, and implementing labels for generative AI.

[5] Specifically, survey respondents from China interpreted this term to mean “made by a human” (as opposed to a naturally occurring phenomenon). However, it is important to note that this discrepant result for the “artificial” label occurred only when this term was presented alone (i.e., without “intelligence”). This suggests that “artificial intelligence” is a term of art with international recognition but may lose much of its meaning when broken up into its component parts.

References

1. S. Ghaffary, What will stop AI from flooding the internet with fake images?, *Vox* (2023). https://www.vox.com/technology/23746060/ai-generative-fake-images-photoshop-google-microsoft-adobe.

2. J. A. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, K. Sedova, Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. arXiv arXiv:2301.04246 [Preprint] (2023). https://doi.org/10.48550/arXiv.2301.04246.

3. T. C. Helmus, "Artificial Intelligence, Deepfakes, and Disinformation: A Primer" (RAND Corporation, 2022); https://doi.org/10.7249/PEA1043-1.

4. Harris Poll, Most Americans support regulating generative AI, *Harris Poll* (2023). https://theharrispoll.com/briefs/regulating-generative-ai/.

5. C. Jackson, M. Newall, B. Mendez, S. Feldman, Americans hold mixed opinions on AI and fear its potential to disrupt society, drive misinformation, *Ipsos* (2023). https://www.ipsos.com/en-us/americans-hold-mixed-opinions-ai-and-fear-its-potential-disrupt-society-drive-misinformation.

6. J. Gottfried, About three-quarters of Americans favor steps to restrict altered videos and images, *Pew Research Center* (2019). https://www.pewresearch.org/short-reads/2019/06/14/about-three-quarters-of-americans-favor-steps-to-restrict-altered-videos-and-images/.

7. H. Farid, Creating, Using, Misusing, and Detecting Deep Fakes. *Journal of Online Trust and Safety* **1** (2022).

8. N. C. Köbis, B. Doležalová, I. Soraperra, Fooled twice: People cannot detect deepfakes but think they can. *iScience* **24**, 103364 (2021).

9. K. T. Mai, S. Bray, T. Davies, L. D. Griffin, Warning: Humans cannot reliably detect speech deepfakes. *PLOS ONE* **18**, e0285333 (2023).

10. S. A. Thompson, T. Hsu, How Easy Is It to Fool A.I.-Detection Tools?, *The New York Times* (2023). https://www.nytimes.com/interactive/2023/06/28/technology/ai-detection-midjourney-stable-diffusion-dalle.html.

11. G. D. Vynck, AI images are getting harder to spot. Google thinks it has a solution., *Washington Post* (2023). https://www.washingtonpost.com/technology/2023/08/29/google-wants-watermark-ai-generated-images-stop-deepfakes/.

12. M. Bennet, Bennet Urges Digital Platforms and AI Developers to Label AI-Generated Content, Stop the Spread of Misinformation, *Michael Bennet* (2023). https://www.bennet.senate.gov/public/index.cfm/2023/6/bennet-urges-digital-platforms-and-ai-developers-to-label-ai-generated-content-stop-the-spread-of-misinformation.

13. C. Goujard, EU wants Google, Facebook to start labeling AI-generated content, *POLITICO* (2023). https://www.politico.eu/article/chatgpt-dalle-google-facebook-microsoft-eu-wants-to-start-labeling-ai-generated-content/.

14. T. Ryan-Mosley, M. Heikkilä, Three things to know about the White House's executive order on AI, *MIT Technology Review* (2023). https://www.technologyreview.com/2023/10/30/1082678/three-things-to-know-about-the-white-houses-executive-order-on-ai/.

15. R. Torres, U.S. Rep. Ritchie Torres Introduces Federal Legislation Requiring Mandatory Disclaimer for Material Generated by Artificial Intelligence, *Ritchie Torres* (2023). https://ritchietorres.house.gov/posts/u-s-rep-ritchie-torres-introduces-federal-legislation-requiring-mandatory-disclaimer-for-material-generated-by-artificial-intelligence.

16. B. Schatz, *AI Labeling Act of 2023* (U.S. Government Publishing Office, 2023; https://www.govinfo.gov/app/details/BILLS-118s2691is) vol. 5.

17. The White House, FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI, *The White House* (2023). https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

18. D. Bartz, K. Hu, OpenAI, Google, others pledge to watermark AI content for safety, White House says, *Reuters* (2023). https://www.reuters.com/technology/openai-google-others-pledge-watermark-ai-content-safety-white-house-2023-07-21/.

19. P. Suciu, 'Created By AI' Warning Labels Are Coming To Social Media, *Forbes* (2023). https://www.forbes.com/sites/petersuciu/2023/08/02/created-by-ai-warning-labels-are-coming-to-social-media/.

20. C. Dunton, Get helpful context with About this image, *Google* (2023). https://blog.google/products/search/about-this-image-google-search/.

21. TikTok, New labels for disclosing AI-generated content, *Newsroom | TikTok* (2023). https://newsroom.tiktok.com/en-us/new-labels-for-disclosing-ai-generated-content.

22. K. Clayton, S. Blair, J. A. Busam, S. Forstner, J. Glance, G. Green, A. Kawata, A. Kovvuri, J. Martin, E. Morgan, M. Sandhu, R. Sang, R. Scholz-Bright, A. T. Welch, A. G. Wolff, A. Zhou, B. Nyhan, Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media. *Polit Behav* **42**, 1073–1095 (2020).

23. T. K. Koch, L. Frischlich, E. Lermer, 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**, 495–507 (2023).

24. G. Pennycook, A. Bear, E. T. Collins, D. G. Rand, The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *Management Science* **66**, 4944–4957 (2020).

25. E. Porter, T. J. Wood, Political Misinformation and Factual Corrections on the Facebook News Feed: Experimental Evidence. *The Journal of Politics* **84**, 1812–1817 (2022).

26. C. Shen, M. Kasra, J. F. O'Brien, Research note: This photograph has been altered: Testing the effectiveness of image forensic labeling on news image credibility. *Harvard Kennedy School Misinformation Review* **2** (2021).

27. Z. Epstein, N. Foppiani, S. Hilgard, S. Sharma, E. Glassman, D. Rand, Do Explanations Increase the Effectiveness of AI-Crowd Generated Fake News Warnings? *Proceedings of the International AAAI Conference on Web and Social Media* **16**, 183–193 (2022).

28. P. Mena, Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook. *Policy & Internet* **12**, 165–183 (2020).

29. W. Yaqub, O. Kakhidze, M. L. Brockman, N. Memon, S. Patil, "Effects of Credibility Indicators on Social Media News Sharing Intent" in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, Honolulu, HI, USA, 2020; https://doi.org/10.1145/3313831.3376213)*CHI '20*, pp. 1–14.

30. C. Martel, D. Rand, Misinformation warning labels are widely effective: A review of warning effects and their moderating features. PsyArXiv [Preprint] (2023). https://doi.org/10.31234/osf.io/48p2a.

31. C. Funk, A. Tyson, B. Kennedy, How Americans view emerging uses of artificial intelligence, including programs to generate text or art, *Pew Research Center* (2023). https://www.pewresearch.org/short-reads/2023/02/22/how-americans-view-emerging-uses-of-artificial-intelligence-including-programs-to-generate-text-or-art/.

32. M. Walker, J. Gottfried, Republicans far more likely than Democrats to say fact-checkers tend to favor one side, *Pew Research Center* (2019). https://www.pewresearch.org/fact-tank/2019/06/27/republicans-far-more-likely-than-democrats-to-say-fact-checkers-tend-to-favor-one-side/.

33. M. Chapman, Google to require disclosure for political ads that use AI to alter imagery or sounds, *PBS NewsHour* (2023). https://www.pbs.org/newshour/politics/google-to-require-disclosure-for-political-ads-that-use-ai-to-alter-imagery-or-sounds.

34. D. Klepper, To help 2024 voters, Meta says it will begin labeling political ads that use AI-generated imagery, *ABC News* (2023). https://abcnews.go.com/Business/wireStory/2024-voters-meta-begin-labeling-political-ads-ai-104714650.

35. D. Klepper, A. Swenson, AI-generated disinformation poses threat of misleading voters in 2024 election, *PBS NewsHour* (2023). https://www.pbs.org/newshour/politics/ai-generated-disinformation-poses-threat-of-misleading-voters-in-2024-election.

36. E. Flitter, S. Cowley, Voice Deepfakes Are Coming for Your Bank Balance, *The New York Times* (2023). https://www.nytimes.com/2023/08/30/business/voice-deepfakes-bank-scams.html.

37. P. Verma, They thought loved ones were calling for help. It was an AI scam., *Washington Post* (2023). https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/.

38. T. Hunter, AI porn is easy to make now. For women, that’s a nightmare., *Washington Post* (2023). https://www.washingtonpost.com/technology/2023/02/13/ai-porn-deepfakes-women-consent/.

39. Meta, How Meta’s third-party fact-checking program works, *Meta for Media* (2021). https://www.facebook.com/formedia/blog/third-party-fact-checking-how-it-works.

40. H. Silverman, Helping Fact-Checkers Identify False Claims Faster, *Meta* (2019). https://about.fb.com/news/2019/12/helping-fact-checkers/.

41. X, About Community Notes on X I X Help, *X Help Center*. https://help.twitter.com/en/using-x/community-notes.

42. Z. Epstein, A. Hertzmann, L. Herman, R. Mahari, M. R. Frank, M. Groh, H. Schroeder, A. Smith, M. Akten, J. Fjeld, H. Farid, N. Leach, A. Pentland, O. Russakovsky, Art and the science of generative AI: A deeper dive. arXiv

arXiv:2306.04141 [cs.AI] [Preprint] (2023). https://doi.org/10.48550/arXiv.2306.04141.

43. T. Weikmann, S. Lecheler, Visual disinformation in a digital age: A literature synthesis and research agenda. *New Media & Society*, 14614448221141648 (2022).

44. B. Paris, J. Donovan, "Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence" (Data & Society, 2019); https://datasociety.net/library/deepfakes-and-cheap-fakes/.

45. N. Schick, Don't underestimate the cheapfake, *MIT Technology Review* (2020). https://www.technologyreview.com/2020/12/22/1015442/cheapfakes-more-political-damage-2020-election-than-deepfakes/.

46. S. Altay, F. Gilardi, Headlines Labeled as AI-Generated Are Less Likely to Be Believed and Shared, Even When True or Human-Generated. PsyArXiv [Preprint] (2023). https://doi.org/10.31234/osf.io/83k9r.

47. C. Longoni, A. Fradkin, L. Cian, G. Pennycook, "News from Generative Artificial Intelligence Is Believed Less" in *2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM, Seoul Republic of Korea, 2022; https://dl.acm.org/doi/10.1145/3531146.3533077), pp. 97–106.

48. Meta, Meta's enforcement of fact-checker ratings, *Meta Business Help Center*. https://www.facebook.com/business/help/297022994952764.

49. Z. Epstein, M. C. Fang, A. A. Arechar, D. Rand, What label should be applied to content produced by generative AI? PsyArXiv [Preprint] (2023). https://doi.org/10.31234/osf.io/v4mfz.

50. C. Martel, J. Allen, G. Pennycook, D. G. Rand, Crowds Can Effectively Identify Misinformation at Scale. *Perspect Psychol Sci*, 17456916231190388 (2023).

51. M. Freeze, M. Baumgartner, P. Bruno, J. R. Gunderson, J. Olin, M. Q. Ross, J. Szafran, Fake Claims of Fake News: Political Misinformation, Warnings, and the Tainted Truth Effect. *Polit Behav* **43**, 1433–1465 (2021).

52. M. Groh, Z. Epstein, C. Firestone, R. Picard, Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences* **119**, e2110013119 (2022).

53. Partnership on AI, PAI's Responsible Practices for Synthetic Media: A Framework for Collective Action, *Partnership on AI - Synthetic Media* (2023). https://syntheticmedia.partnershiponai.org/.

54. L. Rosenthol, “C2PA: the world’s first industry standard for content provenance (Conference Presentation)” in *Applications of Digital Image Processing XLV*, A. G. Tescher, T. Ebrahimi, Eds. (SPIE, San Diego, United States, 2022; https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12226/2632021/C2PA--the-worlds-first-industry-standard-for-content-provenance/10.1117/12.2632021.full), p. 26.

55. D. Citron, R. Chesney, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review* **107**, 1753–1820 (2019).

56. M. Hameleers, F. Marquart, It’s Nothing but a Deepfake! The Effects of Misinformation and Deepfake Labels Delegitimizing an Authentic Political Speech. *International Journal of Communication* **17**, 1–21 (2023).

57. J. Ternovski, J. Kalla, P. Aronow, The Negative Consequences of Informing Voters about Deepfakes: Evidence from Two Survey Experiments. *Journal of Online Trust and Safety* **1** (2022).

58. E. Hoes, B. Aitken, J. Zhang, T. Gackowski, M. Wojcieszak, Prominent Misinformation Interventions Reduce Misperceptions but Increase Skepticism. PsyArXiv [Preprint] (2023). https://doi.org/10.31234/osf.io/zmpdu.

59. J. P. Benway, Banner Blindness: The Irony of Attention Grabbing on the World Wide Web. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* **42**, 463–467 (1998).

60. M. Burke, A. Hornof, E. Nilsen, N. Gorman, High-cost banner blindness: Ads increase perceived workload, hinder visual search, and are forgotten. *ACM Trans. Comput.-Hum. Interact.* **12**, 423–445 (2005).

61. C. Wittenberg, B. M. Tappin, A. J. Berinsky, D. G. Rand, The (minimal) persuasive advantage of political video over text. *Proceedings of the National Academy of Sciences* **118**, e2114388118 (2021).