Written input for the AI Office Working Group 2 on Disclosure of deep fakes and certain AI generated text

By

Prof. Dr. Natali Helberger[1] (University of Amsterdam, Algosoc Program and AI, Media & Democracy Lab), Marilu Miotto[2] (University of Rotterdam, Algosoc Program), Hannes Cools (University of Amsterdam, AI, Media & Democracy Lab)[3][4]

19 November 2025

Informed by academic research, we would like to take the opportunity to submit a written response to the following questions: 2. c - Who is the beneficiary of the information? How is the target audience being taken into account? 2.d - How could transparency be improved. Which practical tools are missing. 2.e – What are best practice examples? 2 - 2. How can we make sure the public understands the disclosures: If disclosures are too subtle (metadata only) or too frequent (spam of "AI-generated" labels) users may ignore them. How should the Code ensure notices/ labels are meaningful without being overwhelming?

The importance of a user-centric perspective

1. Labels and information disclosures need to be informed by a research-driven understanding of users' information needs and the intended and unintended effects of disclosures. What those needs and effects are and how they must inform design and content of information disclosures very much depends on what the goal of Art. 50 AI Act is: to flag that content is synthetic content, to warn, or to empower? Citizen's information needs depend on the actual information goal.

Flagging synthetic content and deepfakes

2. If the goal is simply to flag that a piece of content (whether text, video, picture or audio) is synthetic, clear but also standardised labels could be the way to go. But: research driven design is essential, because research also shows that people tend to overlook labels (here and also here (labels for the context of political advertising)). In other words, ill-designed labels do not help users. Also, labels must not be hidden or difficult to spot.

[1] N.Helberger@uva.nl
[2] miotto@essb.eur.nl
[3] h.cools@uva.nl
[4] This contribution builds on, among others, an overview of transparency research by Hannes Cools, Claes de Vreese, Abdo El Ali, Natali Helberger, Pooja Prajod, Nicolas Mattis, Sophie Morosoli, Laurens Naudts and Teresa Weikmann, Tackling the Transparency Puzzle: Five Perspectives from AI Disclosure Research in News, AI, Media & Democracy Lab, https://www.aim4dem.nl/tackling-the-transparency-puzzle/

3. Also, generic labels whose only message is that a piece of content has been generated by AI can confuse citizens, create the expectation that content has been always written without human supervision, or have negative effects on credibility and signal that all synthetic content is manipulative and deceptive, even where this is not the case. Empirical research shows that labelling news headline as AI generated induces scepticism as audience assume full automation without human supervision. Instead, clarifying the specific role that AI had in the process of creating content, the data that has been used, and what has been done to ensure content is trustworthy gives people actual cues to be able to assess the actual trustworthiness of synthetic content.
4. It is equally important to consider the second-order effects of information interventions. Research (here and here) has shown that exposure to realistic video-based synthetic content (also known as deepfakes) leads to reduced trust in news media in general and reduced perceived credibility of news content. Furthermore, more recent research suggests that deepfake labels increase the salience of deepfakes, thereby having an even more negative effect on "authentic" news, especially on individuals' decisions to share good-quality information.
5. These findings further highlight the importance of the label design. In this context, research on media literacy and realistic video-based synthetic content tested different methods of informing individuals about the synthetic nature of the content, distinguishing between a simple warning, actual tips for detecting this content, and a combination of the two (in different formats). What these studies suggest is that giving actual tips on how to detect synthetic content increases people's ability to detect it, along with confidence in the individual's ability to do so; equally important: this kind of intervention does not seem to increase scepticism towards news media. Also, different kinds of content (audio, text, video) may require different information designs, suggesting that there is not one standard label for all kinds of content.

Enabling citizens to make informed choices

6. Not all synthetic content is false, manipulative or deceptive. There can be many beneficial and legitimate users of GenAI in the production of content, for example to protect privacy, avoid algorithmic censorship, enhance content or offer content in different, fluid formats. Also, deepfakes, unlike traditional forms of disinformation, can be used for other, perfectly legitimate purposes, such as entertainment and artistic reasons.
7. Interestingly, recent research looked at the effect of labelling of video parodies generated with AI, and political satire in particular. Labelling such content actually improved individual comprehension and users' ability to understand the parody, leading to higher enjoyment and a deeper critical assessment. Research findings like these suggest that labelling this kind of content might not be as detrimental as expected.
8. However, research on non-entertainment oriented content also found that simply informing users that content is AI generated does not give them sufficient

information to assess whether a piece of content is trustworthy, accurate and truthful.

9. If the goal is to enable citizens to make informed choices how to value and trust a piece of AI-generated content, they need more information than a simple label, including what the actual implications are if a piece of content has been AI generated or manipulated and why it should be trusted. Group interviews with Dutch audience members have revealed a strong desire for transparency, particularly with an emphasis on what data AI relies on, and how it generally influences the news they consume. Also, audiences expressed the need for source disclosures and transparent internal guidelines. Other suggestions included distinguishing between 'fully AI-generated content', 'partially AI-assisted work', and 'low AI-augmented material. For more concrete suggestions of the kind of information citizens indicated they would find helpful, see here. One potential good practice examples can be found here, and another here.
10. More generally: News media can be valuable pioneers in showing how to consider AI disclosures as a continuum (and not very binary: either disclose or not disclose). This continuum could therefore be an important insight for the working group as well.
11. Much will depend on the right amount of information. One experiment (publication pending) with 3 disclosure conditions, showed that the two-sentence disclosure induced most trust, but the full disclosure condition (5 sentences of text) resulted in less trust.
12. Importantly, in focus group research we also found that users express not only a need for transparency, but also expectations that there is an effective compliance mechanism.

Empowering users

13. If the goal is behind Art. 50 (4) AI Act is, as recital 133 suggests, to safeguard integrity and trust in the information ecosystem, transparency needs to be accompanied by actionable citizen rights that give people some level of control over their information environment.
14. Research demonstrates that transparency alone, without meaningful choices or means of redress, is not enough to instil trust or give people a sense of agency (here and also here). This can include being able to filter news that has been written by an AI or a human; being able to report an article, respectively to complain to a news organisation; and being able to flag biases that they see in news produced by a particular source.
15. For the same reasons, disclosure obligations depend on the role of AI in the production of content. For less impactful AI tasks, such as spell-checking or translation, detailed disclosures may not be necessary. On the contrary, tasks like generating headlines or writing articles that are closer to the generation of actual text with AI, more explicit disclosures are needed. Research concludes that by offering different levels of transparency based on the AI's role in the news production process, news organizations can build trust while addressing readers' concerns. Research also found that comprehensive, layered disclosures can be a

more user-centric form of informing users, while also reducing cognitive load and responding to individual information preferences.

16. The exclusion of editorially controlled content is not aligned with citizens needs. Empirical research showed that for news readers it is important to know whether an article has been written by AI or not, irrespective of whether this has been done under editorial control or not (see also here).

Conclusion

17. Taking into account the information needs of users is critical when designing labels and information disclosures. Therefore, an aspect that the guidelines should also include is the need to invest more resources in such design-specific research and experimentation.
18. Transparency as a principle, and disclosure as a practice is not done with a label, but must be approached as a continuous process of disclosure and communication with the audience.
19. The current transparency requirements in the AI Act are very limited, making users aware that content they see is created with the help of AI. In order to be able to assess synthetic content upon its accuracy, veracity and trustworthiness, users need additional, well-crafted information.
20. However, transparency alone is insufficient to generate trust, in addition effective enforcement and giving users concrete means of exercising agency are critical.