interface

Contribution to the European Commission's consultation on the draft guidelines on transparency obligations under the AI Act

by Nicole Lemke and Lena-Maria Böswald, Senior Policy Researchers at interface

We appreciate the European Commission's efforts to provide guidance on transparency obligations for providers and deployers of AI systems under the AI Act while simultaneously working on the Code of Practice on transparency for AI-generated content. Under Article 50 of the AI Act, providers of AI systems must ensure machine-readable marking and detectability of AI-generated or AI-manipulated content. Deployers must disclose when AI is used to create realistic synthetic content, including deepfakes, by clearly informing users that such content is AI-generated or manipulated. With our contribution, we want to highlight that transparency disclosure as designed in Article 50 is only a necessary, not a sufficient, condition for trust in AI systems. We close by pointing out that labelling and purely personal, non-professional activity does not justify illegality.

On Article 50(1): Transparency for interactive AI systems

We welcome the transparency disclosure for interactive AI systems pursuant to Article 50(1) AI Act. The Commission states that "the purpose of this information obligation is to enable those natural persons to take informed decisions regarding the system's outputs, to avoid that those natural persons over rely on such systems, and to allow those natural persons to calibrate their trust in the content and the interactions accordingly". We point out that transparency disclosure as designed in Article 50(1) is only a necessary, not a sufficient condition for trust in AI systems.

Labelling does, for example, not tell users whether the content is accurate, misleading, deceptive or manipulative; only that content is AI-generated or not. Harmless content receives the same labelling as a potentially misleading deepfake. Moreover, AI-generated content can be persuasive in ways that bypass critical reasoning. This emerges directly from model properties[1] such as (but not limited to) sycophancy, their

[1] El-Sayed, S., Akbulut, C., McCroskery, A., Keeling, G., Kenton, Z., Jalan, Z., Marchal, N., Manzini, A., Shevlane, T., Vallor, S., Susser, D., Franklin, M., Bridgers, S., Law, H., Rahtz, M., Shanahan, M., Tessler,

ability to generate believable responses irrespective of their accuracy, and the false authority with which AI-generated content is often presented to the user. Importantly, disclosing that content is AI-generated does not diminish its persuasive effects.[2] Research shows, for example, that there is no easy fix for mitigating the potentially harmful effects of AI-generated visual disinformation, with no corrective format effectively reducing the perceived credibility of AI-generated visual disinformation or agreement with the false claims it portrayed.[3] Moreover, the "obvious nature" of AI interaction and the expertise of the users interacting with the system, as specified in the exceptions to transparency disclosures, do not protect individuals from these risks.

AI-generated content can thus generate false trust on an individual level both by confidently presenting false or biased information and by subtly influencing users' decision-making without their knowledge. This may happen even in seemingly innocuous situations, such as with biased AI writing assistants.[4] At the societal level, if adopted at scale, this may over time degrade the epistemic quality of our information environment and the knowledge we produce, although the long-term consequences of this remain an active area of inquiry. Transparency disclosure, while a useful first step, does not sufficiently safeguard individuals and society against these risks. We recommend that the guidelines acknowledge the limits of disclosure as a standalone mechanism and signal openness to evidence-based updates as research on AI persuasion and its effects mature.

Article 50(4): Labelling of deep fakes and certain text publications

Article 3(60) AI Act defines deepfakes as "AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful." At the same time, the guidelines explain that, for AI-generated audiovisual content to be considered a deepfake, it suffices for the subjects to "resemble someone or something that can exist or could have existed in reality." In labelling practice, this will mean that AI-generated or manipulated audiovisual content must be clearly labelled if the depicted scenario can be hypothetically real; where content is evidently artistic, creative, satirical, or fictional, only minimal and non-intrusive disclosure is required.

M. H., Douillard, A., Everitt, T., & Brown, S. (2024). *A mechanism-based approach to mitigating harms from persuasive generative AI*. arXiv. https://arxiv.org/abs/2404.15058

[2] Gallegos, I. O., Shani, C., Shi, W., Bianchi, F., Gainsburg, I., Jurafsky, D., & Willer, R. (2026). Labeling messages as AI-generated does not reduce their persuasive effects. *PNAS nexus*, *5*(2), pgag008.

[3] Weikmann, T. E., Tulin, M., Hameleers, M., & de Vreese, C. (2025, April 4). Label sources for AI-generated visual disinformation. https://doi.org/10.17605/OSF.IO/R38AE

[4] Williams-Ceci, S., Jakesch, M., Bhat, A., Kadoma, K., Zalmanson, L., & Naaman, M. (2026). Biased AI writing assistants shift users' attitudes on societal issues. *Science Advances*. https://doi.org/10.1126/sciadv.adw5578

In light of Article 50(4)'s purpose of reducing risks of impersonation, deception, misinformation and manipulation, the guidelines and the Code of Practice on transparency for AI-generated content should keep these risks in mind when using labels. One possibility could be declaring AI-generated content as "altered by AI".

On Article 50(2): Marking and detection of AI-generated or manipulated content

Having explained why transparency is not the only condition for trust in AI systems, we strongly support the assessment that marking or labelling applied to AI-generated or manipulated content should not influence the assessment and the decision on the illegality of the specific content under other regulatory frameworks. As the AI Act does not regulate content as such but regulates AI systems and their use instead, this is where the Digital Services Act and national laws come into play.

On Article 2(10): Purely personal, non-professional activity

Article 2(10) excludes obligations of deployers who are natural persons using AI "in the course of a purely personal, non-professional activity." The exclusion "should [...] not encompass criminal activities since these should not be considered purely personal, even if no economic benefit is sought or attained." Non-consensual sexually explicit deepfakes used to threaten or coerce someone in a private chat are widely recognised as causing serious harm and are increasingly framed, in EU and national debates, as conduct that should be criminalised. We ask the Commission to expressly confirm in their guidelines that criminal and harmful conduct involving non-consensual sexual deepfakes falls outside the scope of Article 2(10), even where no economic benefit is sought or obtained and the abuse occurs in private or semi-private settings (such as direct messages or private chats).