Written contribution by Marilu Miotto[1], Natali Helberger,[2] both Algosoc Research Program

Informed by academic research, we welcome the opportunity to submit written feedback on the First Draft of the Code of Practice on Transparency of AI-Generated Content.

We would first like to highlight the positive elements of the draft. In particular, the presence of a taxonomy that differentiates among uses of generative AI is a valuable and welcome choice, as it increases transparency. Simply flagging content as "AI-generated" may lead users to assume that the content is entirely produced by AI, even when AI played only a partial or assistive role (Wittenberg et al., 2024), resulting in potentially misleading interpretations. By contrast, a differentiated taxonomy can help users better understand the role played by AI in content creation. Furthermore, the inclusion of an enforcement mechanism allowing third parties to flag undetected or unlabelled content is also a positive feature. This approach enables the participation of external actors and may contribute to a more effective application of the regulation.

That said, we have several concerns regarding label design, enforcement mechanisms, and the user-generated content value chain, particularly on social media platforms.

Label Design

How labels are designed is fundamental to the effectiveness of transparency obligations. Existing research on the labelling of disinformation highlights that commonly used warning labels can go unnoticed by users (Dobber et al., 2025; Sharevski & Zeidieh, 2023). Moreover, when labels are noticed, even relatively small design changes can substantially affect how users interpret their intended meaning (Guo et al., 2024). Users can also interpret labels differently, depending on the context (e.g. whether GenAI is used in the context of news or entertainment) (Burrus et al., 2024).

The literature already provides useful insights that could inform the implementation of transparency measures (Clayton et al., 2020; Eslami et al., 2018; Hsueh & Chen, 2025; Obar &

[1] University of Rotterdam, miotto@essb.eur.nl
[2] University of Amsterdam, n.helberger@uva.nl

Oeldorf-Hirsch, 2020; Pennycook et al., 2020; Sharevski et al., 2022; Xiao & Benbasat, 2015). For example, informing users about how to recognise AI-generated content in the future (Xiao & Benbasat, 2015), or explicitly labelling some content as human-generated (Pennycook et al., 2020), may help prevent users from developing excessive scepticism toward online information. Another study found that users prefer detailed and layered disclosures because of more transparency (Gamage et al., 2025; Prajod et al., 2026). Prajod et al. however, also signalled a paradoxical trade-off between trust and transparency, finding that detailed disclosures can also result in lower trust scores. Much will therefore depend on the information provided, and how well it succeeds in signalling not only that a piece of content has been AI generated, but also what has been done to make sure the content is trustworthy (see also below quality labels).

What information do users need

One study found that **disclosing technical details** (e.g. the tools or models used) can increase trust (Chen et al., 2025). The same study found that a significant part of users to the study wanted **clarity on whether an image was fully or partially AI generated**, and another share wanted information on the content's intent (whether to inform, entertain, etc). Information about the **source** (that is who generated the AI manipulated or generated content) was another piece of information that can help users in making an informed choice, as well as the presence of **human oversight** (Venkatraj et al., 2025). These findings are confirmed by Gamage et al. (2025), reporting that users asked for information on the image's source, but also the **AI software used, o**r an explanation of **how or which parts of the image was altered**. Burrus et. Al. found that information **when and how GenAI has been used** (e.g. simple auto-enhancement vs more complex acts of GenAI manipulation) impacted perceptions of authenticity (Burrus et al., 2024).

Providing too much or overly complex information may be counterproductive, as information overload can lead users to ignore warnings altogether (Obar & Oeldorf-Hirsch, 2020). Gamage et al. (2025) identified and tested four design dimensions (Label Sentiment, Icon/Colour, Position, and Label Detail. Relevant is also the study by Kusters et al. (2026) that developed 4 prototypes on disclosing human-AI collaborations, informed by the testing of 69 disclosure designs.

Therefore, a key recommendation is that icons and accompanying messages should be tested with users prior to deployment to assess their visibility, interpretability, and overall efficacy.

Warning vs quality labels

On a similar note, another aspect to consider is that, as currently proposed, the label primarily serves as a warning, without providing information on how AI was used in the content production process. Introducing a more informative or "quality" label (similar to the example illustrated in Figure 4) that embeds (meta)data regarding how AI has been used and the policies/safeguards in place to make sure AI-created or assisted content is sincere could help users to make an informed assessment of the trustworthiness of such content. A quality-label approach would also help to a) address the above-mentioned trust-transparency-trade-off, and b) avoid a situation in which the absence of a label may mislead users into believing that a piece of content has been human generated. This approach finds also support in existing academic research. One empirical (quantitative and qualitative) study found that certification labels can be effective in increasing end-users trust and willingness to use AI and made concrete design suggestions (Scharowski et al., 2023). Again, requiring user testing prior to deployment is essential to avoid risks related to a lack of clarity or information overload (Obar & Oeldorf-Hirsch, 2020) and to account for the fact that users have different information needs in different contexts (e.g. political and entertainment; see Gamage et al. (2025)).

Enforcement

With regard to enforcement, the involvement of third parties is a positive feature, as it enables multiple actors to contribute to achieving the regulation's objectives. However, the absence of academic institutions and researchers from the list of eligible third parties (Measure 2.3) does not appear well justified.

This exclusion is particularly striking given the already recognised benefits of involving academia (alongside civil society and industry) in efforts to counter disinformation and related phenomena. EU policy initiatives themselves acknowledge the value of such multi-stakeholder cooperation, including the contribution of academic expertise (*Cooperating with Fact-Checkers,*

Civil Society, Media and Academia - European Commission, n.d.). Also, transparency to where and how unlabelled content can be report is necessary, also in relation to consumers.

User-Generated Content Value Chain

Finally, the Code of Practice appears to focus primarily on the first uploader of AI-generated content. However, it remains unclear how the framework applies to content that is subsequently reposted, especially across platforms and between AI systems, editing software, and social media platforms.

For example, if an AI-generated image is later edited without the use of AI and then reposted, should it still be labelled as AI-generated? From a transparency perspective, the answer is likely yes. More importantly, this raises the question of whether there are mechanisms to ensure that information about AI involvement (potentially embedded as metadata) is preserved throughout the creation, editing, and posting process.

Responsibility in this respect should not rest solely with users. While obligations could be imposed through the terms and conditions of AI tools, intermediary software providers and platforms should also facilitate the preservation and transmission of such information, so that users without malicious intent are supported in complying with transparency requirements (C2PA, 2022).

Cited Literature

Burrus, O., Curtis, A., & Herman, L. (2024). *Unmasking AI: Informing Authenticity Decisions by Labeling AI-Generated Content*. Interactions. https://doi.org/10.1145/3665321

C2PA. (2022, January 6). *C2PA Releases Specification of World's First Industry Standard for Content Provenance – Coalition for Content Provenance and Authenticity (C2PA)*. https://c2pa.org/c2pa-releases-specification-of-worlds-first-industry-standard-for-content-provenance/

Chen, J., Wang, T.-Y., Williams, M., Jordan, N. A., Shao, M., Zhang, L., & Fussell, S. R. (2025). Examining the Impact of Label Detail and Content Stakes on User Perceptions of AI-Generated Images on Social Media. *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing*, 270–275. https://doi.org/10.1145/3715070.3749237

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

Cooperating with fact-checkers, civil society, media and academia—European Commission. (n.d.). Retrieved 22 January 2026, from https://commission.europa.eu/topics/countering-information-manipulation/cooperating-fact-checkers-civil-society-media-and-academia_en

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

Eslami, M., Krishna Kumaran, S. R., Sandvig, C., & Karahalios, K. (2018). Communicating Algorithmic Process in Online Behavioral Advertising. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3174006

Gamage, D., Sewwandi, D., Zhang, M., & Bandara, A. K. (2025). Labeling Synthetic Content: User Perceptions of Label Designs for AI-Generated Content on Social Media. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–29. https://doi.org/10.1145/3706598.3713171

Guo, C., Guo, Z. 'Sherry', Zheng, N., & Guo, C. 'John'. (2024, January 3). *All Warnings Are Not Equal: A User-Centered Approach to Comparing General and Specific Contextual Warnings against Misinformation*. https://doi.org/10.24251/HICSS.2024.288

Hsueh, C.-H., & Chen, C.-H. (2025). A Study on the Visual Cue Designs in Rich Notification User Interface. In H. Mori & Y. Asahi (Eds), *Human Interface and the Management of Information* (pp. 90–101). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-93819-1_7

Kusters, A., Prajod, P., Cesar, P., & Ali, A. E. (2026). *More Human or More AI? Visualizing Human-AI Collaboration Disclosures in Journalistic News Production* (No. arXiv:2601.11072). arXiv. https://doi.org/10.48550/arXiv.2601.11072

Obar, J. A., & Oeldorf-Hirsch, A. (2020). The biggest lie on the Internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, *23*(1), 128–147. https://doi.org/10.1080/1369118X.2018.1486870

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *Management Science*, *66*(11), 4944–4957. https://doi.org/10.1287/mnsc.2019.3478

Prajod, P., Cools, H., Röggla, T., Venkatraj, K. P., Kusters, A., ElKattan, A., Cesar, P., & Ali, A. E. (2026). *Full Disclosure, Less Trust? How the Level of Detail about AI Use in News Writing Affects Readers' Trust* (No. arXiv:2601.09620). arXiv. https://doi.org/10.48550/arXiv.2601.09620

Scharowski, N., Benk, M., Kühne, S. J., Wettstein, L., & Brühlmann, F. (2023). Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 248–260. https://doi.org/10.1145/3593013.3593994

Sharevski, F., Devine, A., Jachim, P., & Pieroni, E. (2022). Meaningful Context, a Red Flag, or Both? Preferences for Enhanced Misinformation Warnings Among US Twitter Users. *Proceedings of the 2022 European Symposium on Usable Security*, 189–201. https://doi.org/10.1145/3549015.3555671

Sharevski, F., & Zeidieh, A. N. (2023). "I Just Didn't Notice It:" Experiences with Misinformation Warnings on Social Media amongst Users Who Are Low Vision or Blind. *Proceedings of the 2023 New Security Paradigms Workshop*, 17–33. https://doi.org/10.1145/3633500.3633502

Venkatraj, K. P., Morosoli, S., Cools, H., Naudts, L., Vreese, C. H. de, Helberger, N., Cesar, P., & El Ali, A. (2025). *Understanding AI Disclosure Needs for News Production and Journalism*. https://ir.cwi.nl/pub/35915/35915.pdf

Wittenberg, C., Epstein, Z., Berinsky, A. J., & Rand, D. G. (2024). Labeling AI-Generated Content: Promises, Perils, and Future Directions. *An MIT Exploration of Generative AI*. https://doi.org/10.21428/e4baedd9.0319e3a6

Xiao, B., & Benbasat, I. (2015). Designing Warning Messages for Detecting Biased Online Product Recommendations: An Empirical Investigation. *Information Systems Research*, *26*(4), 793–811. https://doi.org/10.1287/isre.2015.0592