caisa Det Nationale Center for AI i Samfundet
The National Center for AI in Society

Transparency of AI-generated content when AI is the norm

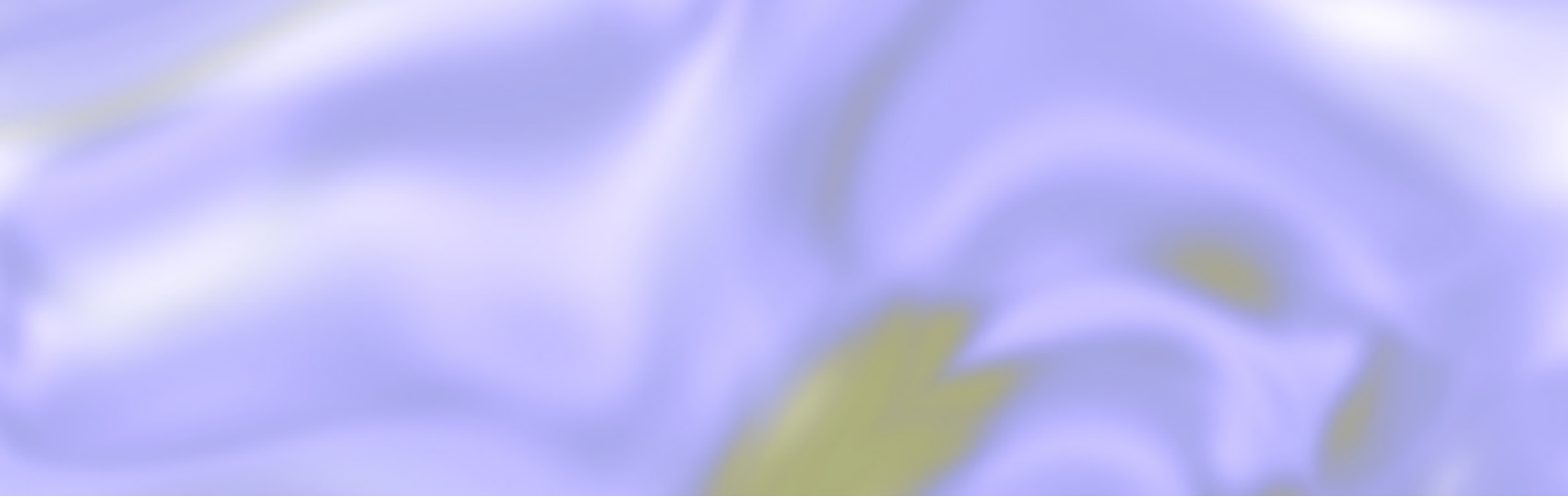
CAISA Research Brief

Authors
Bechmann, Anja; de Vreese, Claes H.; Lewandowsky, Stephan; Helberger, Natali; Søgaard, Anders; Møller, Naja Holten; Shklovski, Irina; Dignum, Virginia; Botan, Madalina; De Gregorio; Giovanni

Editors
Lomborg, Stine; Feldt, Johannes N.

- Publication date: 7 May 2026
- Published by: CAISA, The National Center for AI in Society, Copenhagen and Aalborg, Denmark

- ISSN: 2795-0646
- Document version: Publisher's PDF, registered version
- Citation for published version (APA): Bechmann, A., de Vreese, C. H., Lewandowsky, S., Helberger, N., Søgaard, A., Møller, N. H., Shklovski, I., Dignum, V., Botan, M., & De Gregorio, G. (2026). Transparency of AI-generated content when AI is the norm. *CAISA - Brief*. https://caisa.dk/forskning/governing-ai-generated-content



Transparency of AI-generated content when AI is the norm

By Anja Bechmann, Claes H. de Vreese, Stephan Lewandowsky, Natali Helberger, Anders Søgaard, Naja Holten Møller, Irina Shklovski, Virginia Dignum, Madalina Botan and Giovanni De Gregorio

Summary

Through six interventions from leading European scholars in their field, this research brief examines the challenges of governing AI-generated content in an information environment where such content is rapidly becoming the norm. Drawing on interdisciplinary perspectives, the contributions assess the effectiveness and limitations of emerging AI transparency governance, particularly labelling requirements under the EU AI Act and the forthcoming Code of Practice on marking and labelling of AI-generated content. While transparency labels are normatively important for informing users about content provenance, research suggests that labelling alone is unlikely to mitigate manipulation, restore trust, or empower citizens. The research brief therefore argues for a broader transparency ecosystem that combines labelling with governance infrastructure, organisational accountability, and ongoing research to develop adaptive, evidence-based approaches to AI transparency.

Introduction: Information integrity in democracies when AI is the norm

Anja Bechmann, School of Communication and Culture, Aarhus University, Denmark

To guide stakeholders in the compliance of AI-transparency obligations, the EU released its second draft of the Code of Practice on marking and labelling in March and will publish its final Code in June, shortly before the transparency provisions of the AI Act (Regulation (EU) 2024/1689) take effect on 2 August 2026.

The goal of the transparency obligations of the AI Act is clear: to make it easier for natural persons to know when they are exposed to and interacting with AI-generated or manipulated content. This feeds into a broader democratic principle of strengthening information integrity and the foundation for the 450 million EU citizens to make informed decisions.

What is the EU code of practice on marking and labelling of AI-generated content?

The EU Code of Practice on marking and labelling of AI-generated content is a voluntary, non-legally binding set of commitments and measures which support providers and deployers to comply with the transparency obligations introduced by Article 50 of the AI Act. Instruments such as the Code of Practice are governance tools that can help create more legal certainty and reduce administrative burdens for companies that adopt it by helping navigate the complex and generic requirements of the AI Act.

Although the code does not work in isolation, the Code on of AI-generated content is designed to specifically address the legal text of the AI Act Article 50 sections 2,4 and 5 [see appendix].

According to the AI Act, not all content needs to have a visible interface label. Only content made by professional users will have to be labelled. Personal and non-professional activities do not need to be labelled. Again, it

is also not all types of content either but only deepfakes[1] and AI-generated and manipulated text for the public of public interest where a natural or legal person does not carry human oversight and take editorial responsibility.

In the broad landscape of AI transparency governance and regulation, we solely focus on article 50 (2, 4, 5). This article exists among other articles of the EU AI Act, as well as other EU and national regulatory initiatives, which importantly also shape how AI is governed e.g. AI agentic decision making, etc.

These regulatory developments build on a clear signal from existing research: most studies indicate that citizens are unable to reliably distinguish between AI-produced and human-produced content (Chein et al., 2024; Hove et al., 2024; Köbis et al., 2021). Given the rapid development of AI technologies and their increasing ability to simulate authentic events, objects, and subjects, there is little reason to assume that citizens' ability to discern such differences will improve over time.

At the same time, industry reports (Graphite.io, n.d.) conclude that AI-written content surpassed human-created content online in 2025, while research increasingly examines the implications of this shift across multiple fields (e.g., Haider et al., 2024). This suggests a fundamental transformation of the information environment: AI-generated content becomes the norm rather than the exception.

This transformation may also reinforce itself. As AI-generated content grows in volume, models may increasingly be trained on datasets containing large proportions of synthetic material, potentially amplifying the detachment between informational outputs and empirical reference points (see also Hicks et al., 2024). While AI clearly offers substantial benefits for societies and individuals (from accelerating medical research to reducing repetitive labour) these developments also contribute to an information environment in which knowledge production changes epistemically and to some extent at least spiral into self-referential synthetic systems that permeate all domains of online and physical world communication settings. This creates several challenges for democratic societies. I highlight two in particular.

First, how can trust in information and institutions (both foundational elements of democratic systems) be maintained in a context where content can be easily fabricated or manipulated to attract attention, advance particular claims, or obstruct public debate through mechanisms such as the liars' dividend, where actors claim true information to be false (Schiff et al., 2025)?

Second, if we cannot tell when, whether, and how AI-generated and manipulated content is based in real-world events, subjects, and objects, how should this be communicated to the citizens? In particular, how can labelling practices remain meaningful in a context where the majority of content is AI-generated or manipulated but AI is used in many ways and for different purposes?

While existing research provides relatively consistent evidence that citizens struggle to distinguish between AI-generated and human-generated content, there is considerably less consensus regarding appropriate policy responses and labelling strategies (see e.g., Altay & Gilardi, 2024; Burrus et al., 2024; Chen et al., 2025; Gamage et al., 2025; Sharevski & Zeidieh, 2023). Some studies suggest the emergence of label blindness, while others find that explicitly mentioning AI improves recognition. Some advocate providing more contextual information, others recommend minimal labels to avoid cognitive overload. Similarly, some studies identify declining trust in labelled content, whereas others observe stable trust levels over time. Although such behaviour seems contradictory, it might be because multiple dynamics operate simultaneously within complex information environments (Bechmann, 2026).

In light of these dynamics, as well as the ongoing development of the Code of Practice and the implementation of the AI Act, the six interventions in this research brief address the dilemmas surrounding AI transparency and labelling and consider how governance frameworks might respond effectively to the changing structure of the information environment. Together, the contributions reflect on how current challenges might best be framed, which dilemmas deserve particular attention, and which policy pathways may offer promising directions for future governance.

[1] Deepfake refers to "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" (AI Act 3 (60).

To label or not: that's not the question

Claes H. de Vreese, Amsterdam School of Communication Research, University of Amsterdam, The Netherlands

Article 50(4) of the EU AI Act is clear: "Deployers of an AI system that generates or manipulates image, audio or video content constituting a deep fake, shall *disclose* that the content has been artificially generated or manipulated" (italics added). But the devil is in the detail. Critics will promptly say that labeling is hard and labeling does not always work. Which is true. But let's stay focused on the ball: The real issue is not whether labelling is necessary, but how to do it effectively in a complex and quickly evolving (AI-driven) information ecosystem.

Labelling serves a fundamental democratic function. It provides transparency, enabling citizens and consumers to distinguish between human-created and AI-generated content. In an era where synthetic content can easily mimic reality, this distinction matters for trust, accountability, and informed decision-making.

Evidence shows that labels are not a silver bullet. It is a persistent challenge: citizens do not always notice labels, and even when they do, they may not act on them. This mirrors findings from other domains, such as nutrition labels or privacy notices, where information is available but often ignored. In other words, labelling can be sub-optimal in practice! Also in journalism, the 'AI labeling' discussion is heated (Cools et al., 2025) and labeling can even lead to less trust (Epstein et al., 2023) though this finding may not apply to all situations (Gallegos et al., 2026). That said, labels and for example transparency information can aid citizens (Binder et al., 2022) and citizens also sometimes accept mixed content (such as advertorial content, see van Reijmersdal et al., 2005).

Either way, an easy-to-recognize AI indicator could be complemented by more detailed information for those who seek it. Such a system respects the reality of user behavior: most people will only engage superficially, very briefly, and on a small to modest size screen. Only a smaller group will want deeper insights. Both having the information available is important.

The European Union is uniquely positioned to establish a single, harmonized framework that applies across member states and platforms. A unified approach would not only enhance clarity but also reduce compliance burdens.

Critics often argue that the challenges of labelling make it impractical. But **not** trying is effectively the same as giving up. AI development and deployment are ongoing processes, and labelling must evolve alongside them. It is a moving target: AI can be embedded in countless systems, from news production to customer service, making it difficult to define when something should be labelled. However, this challenge will persist in one, two, or five years. Delaying action only postpones learning.

Instead, policymakers should embrace trial and error. Labelling systems should be tested, evaluated, and refined drawing on systematic research. Continuous adaptation is not a weakness but a necessity in a rapidly changing technological landscape.

Finally, effective labelling requires more than good design. It demands a robust framework for implementation, monitoring, compliance, and enforcement. Without these, even the best-designed system will fail in practice.

The question, then, is not whether to label AI content. That debate is settled. The real challenge lies in labeling well.

Is transparency enough to deal with a "post referential world"?

Stephan Lewandowsky, School of Psychological Science University of Bristol, United Kingdom

As Professor Bechmann notes in her introduction, at a time when AI-generated content is on track to dominate the information ecosystem—ultimately feeding back into AI training—we risk entering a "post-referential world" where information is no longer necessarily anchored to real-world events but instead is bounded primarily by what machines and humans can collectively imagine. This post-referential world seems incompatible with known forms of societal governance, in particular democracy, because it permits neither accountability nor evidence-based decision-making. Preventing this world from emerging is therefore one of the defining—and most formidable—tasks of the 21st century.

The EU is at the forefront of online regulatory efforts and the only organization worldwide that may have the power and political will to reign in unfettered technological development (for a review and recommendations, see Scharfbillig, et al., 2026). The second draft of the EU's *Code of Practice on Transparency of AI-Generated Content* that was released in March, 2026 (European Commision, 2026), presents one further step in that direction and covers transparency obligations for providers and deployers of generative AI systems under Article 50 of the AI Act, which requires that AI-generated content be marked, detectable, and disclosed to users before or at first exposure. The premise underlying transparency measures is that citizens will be shielded from unwarranted manipulation if they are aware of the provenance of information (e.g., Dobber et al., 2023).

Available behavioral research provides only partial support for this premise. Several lines of research have identified the limits of transparency.

There has been much concern about the potential manipulative power of so-called "microtargeting", that is the use of messages that are targeted at a recipient using psychological information such as personality. Recent evidence has confirmed that AI-generated tailoring based on users' personality can boost the persuasive power of political ads (Simchon et al., 2024). Carrella, Simchon, Edwards, and Lewandowsky (2025) tested whether warning users that they were being microtargeted would eliminate this persuasive advantage of microtargeting. Across three within-subject studies (total $N \approx 1{,}767$), personality-targeted ads were consistently rated as more persuasive than non-targeted ones, replicating the core targeting effect. Critically, warning popups explicitly alerting participants to the microtargeting had no meaningful impact — equivalence tests confirmed the warning effect was too small to matter in practice.

In another line of research, Clark and Lewandowsky (2026) investigated whether transparency warnings could neutralise the influence of AI-generated deepfake videos. Participants in three preregistered between-subjects experiments (N = 175, 275, 223) watched a deepfake video of a fictional character appearing to confess to a crime or moral transgression. In some conditions the video was preceded by a specific warning that it had been identified as a deepfake. Across all three experiments, participants who watched the deepfake video rated the target as significantly more guilty than those in a control condition who had not been exposed to relevant information — and critically, this influence persisted even when participants had been explicitly warned beforehand and even among the subset who indicated they believed the warning and knew the video was fake. The persistent effect of content on people's judgments even after it has been identified as fake again points to the limits of transparency as a shield against manipulation.

In summary, transparency measures such as labelling and disclosure are a normative good — citizens have a right to know the provenance of the information they encounter, and this right alone justifies the EU's Code of Practice. However, the behavioural evidence reviewed here suggests that transparency is unlikely to be sufficient as a standalone shield against manipulation in a post-referential world.

Informed but not empowered

Natali Helberger, Institute of Information Law, University of Amsterdam, The Netherlands

The AI transparency labelling obligations in Art. 50 (4) AI Act are one of the few end user-facing provisions in the AI Act, and an important step forward in helping users to navigate an increasingly synthetic world. Without knowing whether a piece of content has been AI generated or not, users do not even have a chance to make an informed decision how to engage. Some form of transparency also responds to the wishes of users who otherwise may feel manipulated if not informed (Piasecki et al., 2025). Insofar, the Code of Practice as an initiative to operationalise these rules is an important step forward towards improving the position of end users.

The core question, however, is what information users need to make an informed decision about a piece of (synthetic) content. Simple transparency labels can inform that a piece of content has been generated by or with the help of AI, but their message is ambiguous: what part has been generated with AI? How was AI used and why? What has, if anything, been done to ensure the accuracy and integrity of that piece of content? And: is the content trustworthy? There is a risk that simple AI labels inform but do little to empower. In the worst case, they have backfire effects and repercussions for trust in the public information ecosystem and for the credibility of trustworthy content (Altay & Gilardi, 2024; Miotto et al., 2025). In line with the objectives of the AI Act, the "AI" label that is proposed in the Code of Practice must warn users of content that is deceptive, manipulative and fake. Not all AI generated content is deceptive, manipulative or fake, however. Generative AI can and will be used for all kinds of legitimate purposes: to illustrate, inform or make information more accessible (Caswell, 2024; Dodds et al., 2026; Jones & Jones, 2019), to exercise freedom of expression rights and evade automated censorship (Philips & Torres, 2024) or to lower the threshold for less well-endowed political parties to engage voters (Helberger et al., forthcoming). If the goal is to enable users to make informed choices, they need additional information that enables them to assess the accuracy, salience and trustworthiness of synthetic content. In addition, and less important, they need affordances that put them in a position to make meaningful choices (Helberger et al., 2025).

All labelling initiatives will not be able to address another potential backfire effect of labelling: helping users understand that the absence of a label does not automatically mean that content is human made. Therefore, an alternative to warning labels that is

worthwhile exploring are voluntary more informative or "quality" signals that help users understand how the creator of such content ensured it is authentic and trustworthy (whether synthetic or not). Such certification labels can be effective in increasing end-users and informed choices (Scharevski & Zeidieh, 2023).

As this brief discussion demonstrates, the transparency obligations in Article 50 AI Act and the Code of Practice are important steps into the right direction but organisations that want to use GenAI *and* earn users trust are well advised to go further.

Two dillemmas of governing in practice:

1. Marking manipulation and communicative equality

Anders Søgaard, Department of Computer Science, University of Copenhagen, Denmark

EU's Code of Practice on Transparency of AI-Generated Content defines a set of rules for marking and detection of AI-generated and manipulated content (Code of Practice section 1, Art. 50(2)). The Code of Practice is motivated by the need for trust in the information ecosystem. The Code of Practice is supposed to ensure that AI system outputs are marked in a machine-readable format and detectable as artificially generated or manipulated, and the layered approach taken to ensuring this relies on metadata, watermarks, fingerprinting, and direct logging. The AI systems that the code of practice is meant to regulate, extract semantics from encoded input from which they decode their output. Such systems make information more widely accessible, e.g., across languages, registers, and data modalities – and can mitigate inequalities around access to information and the ability to communicate to a wider audience.

One cautious response to the code of practice is whether marking of artificially manipulated outputs may be at odds with equal access to large-scale communication (communicative equality)?

Dyslexics rely on AI assistance to write parliamentary speeches, op-eds, literary works, or scientific papers, but outputs marked as AI-manipulated may easily be considered second class. For texts, it will be extremely difficult to detect manipulation unless relying on direct logging. If providers are forced to mark dyslexic people's texts, we may reintroduce an inequality that we were planning to eradicate, e.g. by reducing the social status of content from specific groups and disproportionately flagging such text when sorting through content with screening procedures. AI-generated text detection can be evaded in the absence of watermarking (Perkins et al., 2024), and since text manipulation introduces a continuum from grammar checking to significant rewriting, watermarking of manipulation is generally impossible. Tools can, on the other hand, be used to authenticate content or verify that content is authentic, including texts written by dyslexics. Modern cameras come with authentication techniques that add meta-data to photographs (Kang et al., 2022). In the same way, keystroke logs can be used to authenticate texts. Developing editors that prevent copy-pasting blocks of text in should be straightforward. We may also want to develop inherently authenticating writing tools, e.g. typewriter-like devices introducing new encrypted file formats. In general, there seem to be reasons to consider flipping the code of practice around.

2. Should automating politeness be disclosed? The case of GPAI content production in clerical practice

Naja Holten Møller and Irina Shklovski, Department of Computer Science, University of Copenhagen, Denmark

Article 50 of the AI Act specifies that only content that constitutes "a deep fake" or "is published with the purpose of informing the public on matters of public interest" must clearly signal its origins. The Code of Practice attempts to give concreteness to these abstract demands, with a clear understanding that in the digital infosphere, content that might have been created with one purpose in mind often slips the boundaries, and there is no putting the genie back in the bottle after. If only the general-purpose AI(GPAI)-generated content that is clearly intended "in the public interest" is marked, the whole exercise can become a fool's errand. Information, after all, can become of public interest long after it was generated and has spread. What are people to think, if some GPAI content they encounter isn't marked and some is? How are they to know why different contexts matter?

As the use of GPAI becomes integrated across myriad domains, we cannot predict how social norms for making sense of such content will evolve, as both the technology and broader social systems are rapidly changing under so many pressures at the same time. The Code of Practice is an attempt to steer this process by providing a set of commitments that would, hopefully, result in a better kind of future where GPAI generated content is usefully integrated, but it does not apply to all contexts by definition. Lucy Suchman (2002) describes artful integration as the situated work through which new technologies are fitted into existing social and

organizational arrangements, not simply installed as neutral tools. Integration, in this sense, directs attention to how people continue to adapt technology to established norms of responsibility, competence, and care.

Consider the following example: In Danish hospitals, GPAI is being tested in patient communication systems through a pilot study to assist in drafting replies to patients' text messages (Chreiteh, 2024). These exchanges are often mundane, enabling the hospital staff (e.g. clerical workers) to offload the additional work of crafting polite replies to GPAI, when, for example, confirming appointment details. However, they are also consequential because they carry professional authority and institutional responsibility. In our case, we observed a reluctance to inform patients about this use of GPAI for content generation, as such disclosure might make patients feel more insecure, potentially undermining the authority of the replies.

Under the EU AI Act Article 50(4) no disclosure is needed in this case, since the clerical workers only communicate to a patient and not a public. On top of that the hospital workers review and edit the replies, filling in the substantive details of upcoming treatments or appointment locations, thus qualifying under the “editorial control” exception. Despite the letter of the law, what role might the *spirit of transparency* play here? Should the patients be informed that the apparent care in the messages is automated? In this context, is it important to clearly label which parts are produced by GPAI and which are carefully added and checked by the hospital staff? What may such labels do to the relationship between hospital staff and their patients? What sort of work conditions are we allowing or creating where the pressure to do more - and more quickly - delegates politeness and care to automated systems?

The Code of Practice does not help us answer these questions, but it does suggest they ought to be asked. Labelling when and where GPAI is used to generate content is an important move, but it does not necessarily imply a hierarchy of value and importance to human vs. AI-generated content. Instead, we must ask what the implications of artful integration for social systems are that automate signals of care, while maintaining control over content that requires oversight and where mistakes can clearly cause harm.

Beyond the label: transparency as practice in the governance of AI-generated content

Virginia Dignum, Department of Computing Science, Umeå University, Sweden

The EU's second draft Code of Practice on marking and labelling of AI-generated content represents a significant step in operationalising Article 50 of the AI Act due to take force in August 2026 (Regulation (EU) 2024/1689). Its streamlined, two-layered approach combining secured metadata with watermarking, and its flexible regime for deployers, reflects important lessons learned from the first draft consultation. Yet labelling and marking, however technically well-designed, address only the surface of a deeper governance challenge.

Bechmann's introduction rightly frames the core democratic dilemma: when AI-generated content becomes the norm rather than the exception, labelling alone cannot restore the epistemic foundations that democratic discourse requires. Labels can shift stated trust when noticed and processed, but empirical evidence on their effectiveness is mixed at best. Studies consistently find that AI disclosure labels have no significant effect on perceived accuracy or sharing intention for misinformation (Liao et al., 2025), that effects are highly context-dependent and contingent on disclosure design (Altay & Gilardi, 2024), and that habituation reduces their impact as AI content becomes ubiquitous. When citizens cannot distinguish AI from human-produced content, labelling alone is unlikely to restore that ability.

The deeper challenge is to move beyond transparency as a regulatory checkbox toward transparency as an organisational and societal practice (Dignum, 2026). This requires distinguishing between different layers of transparency. Data transparency concerns the sources, composition, and known limitations of training data. Logic transparency addresses the objectives, decision rules, and optimisation criteria of the system. Risk transparency communicates the potential harms, failure modes, and contextual limitations to those affected (Dignum, 2025). Knowing that content is AI-generated is not the same as knowing how it was produced, on what data, for what purpose, or with what likely effect on the reader. Each of these layers requires different mechanisms, different audiences, and different accountability structures. A label addresses only the first, and only partially at that.

This matters because transparency is not simply an information problem but a power relation. Citizens, regulators, and civil society cannot meaningfully contest or evaluate AI-generated content without access to the logic and risk layers. Disclosure of provenance is a starting

point, not an end point. Treating it as sufficient risks creating the appearance of accountability without its substance.

For the Code of Practice to achieve its broader transparency objectives, it must therefore be understood not as an end point but as infrastructure for a broader transparency ecosystem. This includes interoperability with audit and impact assessment obligations under the AI Act, meaningful stakeholder engagement, and governance capacity inside the organisations deploying these systems.

Labelling tells citizens that AI was involved. Governance determines whether that information is meaningful, contestable, and consequential. The icon matters. The organisational and regulatory infrastructure behind it matters more.

Implications: From surface level labels to transparency ecosystems

Anja Bechmann

Madalina Botan, College of Communication and Public Relations, National University of Political Studies and Public Administration, Romania

Giovanni De Gregorio, Católica Global School of Law and Católica Lisbon School of Law, Portugal

While the research-based interventions in this research brief overall support the AI Act's focus on AI transparency, they clearly outline limitations, dilemmas, and challenges to labelling and marking as standalone transparency methods.

The legislative intent suggests that labelling can be effective, but the contributors here present a more sceptical perspective. If 'effectiveness' is defined as preventing manipulation or restoring trust, current labelling policies may be ineffective. As Levandowsky rightly shows in his studies, warning labels show no significant effect on belief or behaviour, and even when warned of deepfakes, users remain susceptible to persuasion. However, as Lewandowsky notes, even if transparency labels alone are not sufficient to shield against manipulation, they remain a normative good, i.e. citizens have a right to know they are interacting with AI, even if it does not change their behaviour. This points to a deeper level of interventions needed for effective AI transparency - an epistemic level where AI does not just mediate information, it alters even the foundations of how we perceive authenticity and authority in AI-permeated landscapes.

To address these limitations, legislators, technology companies, researchers and civil society representatives as a collective "we" must move from 'labels' to 'ecosystems.' This is in line with Dignum's intervention, which highlights that transparency is a power relationship that goes beyond labelling and is mediated by organisational and societal practices in which AI deployers' governance capacities and the broader regulatory infrastructure matter more. Thus, the EU Code of Practice on AI transparency should not be treated as "an end point", but as the governance "infrastructure of a broader ecosystem" where organizational accountability matters more than the icons or labels on the screen.

In a world where AI content is becoming the norm, the nature of the labelling itself changes. As Bechmann discusses, if the majority of content is AI-assisted, labelling with a broad interpretation of deepfake risks becoming noise. In this context, effective labelling might mean to clarify the AI and human involvement in the content and certifying the origin of the content through provenance authentication rather than just tagging the AI.

The integration of AI into specific professional norms shows further complexities. A transparency ecosystem for AI permeates potentially sensitive contexts of life. Møller and Shklovski show that in healthcare, even if a label is not mandatory by law; it can create a moral dilemma where not labelling medical messages might undermine patient trust and affect the doctor-patient relationship if it is not clear what part of the communication has been assisted by AI and what has been carefully checked by hospital staff. Furthermore, Søgaard introduces a critical risk of discriminatory transparency – marking AI-assisted text with machine-readable metadata could stigmatize individuals with disabilities, such as dyslexics, who use AI as accessibility tools. This is a real-life context where the social cost of transparency may outweigh its benefits. Watermarking and AI-metadata might provide algorithms with the wrong signal instead of empowering people. Søgaard, therefore, suggests additionally to flip the Code around and label human made content instead.

Helberger explicitly challenges the ambiguity of current labels by pointing out that labels do little to empower citizens unless they deliver more details on what has been manipulated and what has been done to ensure accuracy. A generic 'AI' tag fails to detail which part of the content was manipulated or whether the output is trustworthy. This ambiguity can trigger a general mistrust across all content, human or synthetic. She suggests that we require additional "quality signals" or certifications that provide details on accuracy and intent rather than binary tags, i.e.

AI/non-AI. Organisations that want "to earn users' trust" are advised, therefore, "to go further" to ensure transparency.

On this basis, the next steps for AI-transparency governance should be to recognize transparency not as a destination, but as infrastructure for the broader information ecosystem. This involves the continuous development of the interface labelling that provides the right information to empower people, integrated with provenance information in meta-data. As de Vreese explains: "The question, then, is not whether to label AI content. That debate is settled. The real challenge lies in labelling well." This requires an iterative process of testing and refinements of disclosure that keeps pace with technical realities rather than a 'one-size-fits-all' solution.

By integrating marking and labelling standards with provenance metadata and organizational accountability, we can move toward an infrastructure that empowers citizens rather than just informing them. The relevance of this approach extends beyond the AI Act framework, as it interacts with the broader regulatory regimes, from public law related to information transparency to further rules of the AI Act targeting high-risk AI systems intended to influence democratic elections. Crucially, the Code of Practice interacts with for instance the Digital Services Act (DSA), especially for Very Large Online Platforms (VLOPs) hosting and disseminating synthetic content at scale. In this layered model, the AI Act governs upstream transparency (disclosure), while the DSA manages downstream implications, such as systemic risk mitigation under Article 34. Even when content is properly labelled, platforms may still need to take additional measures if that content threatens fundamental rights or civic discourse.

It is important to underline that the transparency marking obligation under the AI Act is primarily aimed at disclosure, ensuring that users are aware when they are interacting with AI-generated content and thereby reducing risks of deception at the point of consumption. Indeed, the purpose is not to provide information on the trustworthiness of the content or make users aware of the sharing of disinformation, which could be relevant for the purposes of the DSA. This research brief has highlighted that research in is not only a supporting method to reach transparency, marking and labelling practices. Integrating research into the transparency governance infrastructure of the broader ecosystem will further develop technical free to use systems and increase our understanding on which label solutions work best in which contexts on what kind of topics for whom. This will heighten the focus on AI transparency not as an end point but as a continuous development build on evidence and innovation across stakeholders. The research contributions of this research brief act as a necessary reality check for the EU AI Act and the Code, calling for a much-needed harmonization of the policy promises made by legislators with the nuanced, often unintended, consequences of digital transparency across the entire information ecosystem.

Acknowledgements

We thank Henrik Palmer Olsen and Anna Rogers for insightful and constructive reviews and Stine Lomborg and Johannes Nissen Feldt for careful editorial feedback that have helped strengthen the manuscript before publication. ChatGPT 5.3 has been used to improve the English language of the introduction and to optimize the summary for a broader audience.

About the authors

Anja Bechmann is Professor at Department of Media and Journalism at the School of Communication and Culture and director of the interdisciplinary DATALAB – Center Digital Social Research at Aarhus University and and fellow at CAISA, The National Center for AI in Society. She is the Chair of the EU Code of Practice on Transparency of AI-Generated Content, AI ACT Article 50(4).

Claes H. de Vreese is Distinguished University Professor of AI and Society at The Amsterdam School of Communication, University of Amsterdam, co-founder of the AI, Media and Democracy Lab, and Professor and Director of the Digital Democracy Centre at University of Southern Denmark.

Stephan Lewandowsky is Professor of Cognitive Science at the School of Psychological Sciences, University of Bristol.

Natali Helberger is Distinguished University Professor of Law and Digital Technology at the Institute for Information Law, University of Amsterdam, and co-founder of the AI, Media and Democracy Lab.

Anders Søgaard is Professor of Natural Language Processing and Machine Learning at the Department of Computer Science at the University of Copenhagen and Chief Scientist at CAISA, The National Center for AI in Society.

Naja Holten Møller is Associate Professor of Human-Centred Computing at the Department of Computer Science at the University of Copenhagen, and Senior Scientist at CAISA, The National Center for AI in Society.

Irina Shklovski is Professor of Communication and Computing at the Department of Computer Science at the

University of Copenhagen and WASP-HS Guest Professor in the Department of Thematic Studies, Gender Studies, Linköping University.

Virginia Dignum is Professor of Responsible Artificial Intelligence at the Department of Computing Science, Umeå University, Sweden.

Madalina Botan is Associate Professor of Media Research Studies and Political Communication at the National University of Political Studies and Public Administration (SNSPA) and a senior researcher specializing in digital media effects and strategic communication. She serves as the Vice-Chair of the EU Code of Practice on Transparency of AI (AI Act Article 50(4).

Giovanni De Gregorio is the PLMJ chair of Law and Technology at Católica Global School of Law and Católica Lisbon School of Law. He serves as the Vice-Chair of the EU Code of Practice on Transparency of AI (AI Act Article 50(4).

About CAISA

The National Center for Artificial Intelligence in Society (CAISA) is a national consortium that gathers researchers from the University of Copenhagen, Aalborg University, Aarhus University, the IT University of Copenhagen and the Technical University of Denmark in close collaboration with the Pioneer Centre for Artificial Intelligence (P1).

As Denmark's independent research center for artificial intelligence in society, CAISA centers the citizen. We carry out groundbreaking interdisciplinary research, and we deliver overview of the most recent scientific breakthroughs. Based on new and interdisciplinary research, we advise decision-makers in the public and private sectors on how they may develop and use artificial intelligence practically, such that it contributes to growth, supports democracy and strengthens digital autonomy.

About CAISA's briefs

CAISA briefs are part of CAISA's effort to ensure that knowledge and novel insights from within the research community come to empower decision-makers in public and private sectors, and, in turn, society-writ-large when faced with the opportunities and risks posed by rapid technological change. CAISA publishes two types of briefs:

Research briefs present research and evidence-based knowledge about AI and society in an accessible way.

Position briefs express the authors' research based and informed assessment of important challenges related to AI and society.

CAISA's briefs are edited by Anders Søgaard, Professor at the Department of Computer Science at the University of Copenhagen and Chief Scientist at CAISA, and Johannes N. Feldt, Research Assistant at CAISA. All briefs are read by – and receive comments from – at least one external independent researcher before publication.

The authors are responsible for the contents of a CAISA brief

Appendix

The legal text from the AI Act, Article 50(2, 4, 5) that constitutes the foundation for the EU Code of Practice on AI marking and labelling (European Commision, 2026):

2. Providers of AI systems, including general-purpose AI systems, generating synthetic audio, image, video or text content, shall ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated. Providers shall ensure their technical solutions are effective, interoperable, robust and reliable as far as this is technically feasible, taking into account the specificities and limitations of various types of content, the costs of implementation and the generally acknowledged state of the art, as may be reflected in relevant technical standards. This obligation shall not apply to the extent the AI systems perform an assistive function for standard editing or do not substantially alter the input data provided by the deployer or the semantics thereof, or where authorised by law to detect, prevent, investigate or prosecute criminal offences.

4. Deployers of an AI system that generates or manipulates image, audio or video content constituting a deep fake, shall disclose that the content has been artificially generated or manipulated. This obligation shall not apply where the use is authorised by law to detect, prevent, investigate or prosecute criminal offence. Where the content forms part of an evidently artistic, creative, satirical, fictional or analogous work or programme, the transparency obligations set out in this paragraph are limited to disclosure of the existence of such generated or manipulated content in an appropriate manner that does not hamper the display or enjoyment of the work.

Deployers of an AI system that generates or manipulates text which is published with the purpose of informing the public on matters of public interest shall disclose that the text has been artificially generated or manipulated. This obligation shall not apply where the use is authorised by law to detect, prevent, investigate or prosecute criminal offences or where the AI-generated content has undergone a process of human review or editorial control and where a natural or legal person holds editorial responsibility for the publication of the content.

5. The information referred to in paragraphs 1 to 4 shall be provided to the natural persons concerned in a clear and distinguishable manner at the latest at the time of the first interaction or exposure. The information shall conform to the applicable accessibility requirements.

References

- 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), pgae403. https://doi.org/10.1093/pnasnexus/pgae403
- Bechmann, A. (2026). Platform Collective Behavior as Democratic Infrastructure: Beyond Causal Effects in the DSA Era, *Journal of Political Communication*. Accepted article.
- Burrus, O., Curtis, A., & Herman, L. (2024). Unmasking AI: Informing Authenticity Decisions by Labeling AI-Generated Content. *Interactions*, *31*(4), 38–42. https://doi.org/10.1145/3665321
- Binder A, Stubenvoll M, Hirsch M, et al. (2022) Why am I getting this ad? How the degree of targeting disclosures and political fit affect persuasion knowledge, party evaluation, and online privacy behaviors. Journal of Advertising 51(2): 206–222.
- Carrella, F., Simchon, A., Edwards, M., & Lewandowsky, S. (2025). Warning people that they are being microtargeted fails to eliminate persuasive advantage. *Communications Psychology*, *3*. https://doi.org/10.1038/s44271-025-00188-8.
- Caswell, D. (2024). Audiences, automation, and AI: From structured news to language models. *AI Magazine*, *45*(2), 174–186. https://doi.org/10.1002/aaai.12168
- Chein, J. M., Martinez, S. A., & Barone, A. R. (2024). Human intelligence can safeguard against artificial intelligence: Individual differences in the discernment of human from AI texts. *Scientific Reports*, *14*(1), 25989. https://doi.org/10.1038/s41598-024-76218-y
- 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, CSCW Companion '25*, 270–275. https://doi.org/10.1145/3715070.3749237
- Clark, S., & Lewandowsky, S. (2026). The continued influence of AI-generated deepfake videos despite transparency warnings. *Communications Psychology*. https://doi.org/10.1038/s44271-025-00381-9.
- Cools, H., de Vreese, C., Ali, A. E., Helberger, N., Prajod, P., Mattis, N., Morosoli, S., Naudts, L., & Weikmann, T. (2025). *Tackling the Transparency Puzzle: Five Perspectives from AI Disclosure Research in News – AI, Media & Democracy Lab*. AI, Media, and Democracy Lab. https://www.aim4dem.nl/tackling-the-transparency-puzzle/
- Dignum, V. (2025). https://aijourn.com/putting-algorithmic-transparency-into-practice/
- Dignum, V. (2026). *The AI paradox*. Princeton University Press.
- Dobber, T., Kruikemeier, S., Helberger, N., & Goodman, E. (2023). Shielding citizens? Understanding the impact of political advertisement transparency information. *New Media & Society*, *26*, 6715–6735. https://doi.org/10.1177/14614448231157640.
- Dodds, T., Zamith, R., & Lewis, S. C. (2026). The AI turn in journalism: Disruption, adaptation, and democratic futures. *Journalism*, *27*(3), 530–544. https://doi.org/10.1177/14648849251343518
- Epstein, Z., Fang, M. C., Arechar, A. A., & Rand, D. G. (2023). What label should be applied to content produced by generative AI? *PsyArXiv*. Europe PMC Preprints (PPR699709). https://doi.org/10.31234/osf.io/v4mfz
- European Commission. (2026, March 5). *Commission publishes second draft of Code of Practice on marking and labelling of AI-generated content*. https://digital-strategy.ec.europa.eu/en/library/commission-publishes-second-draft-code-practice-marking-and-labelling-ai-generated-content
- 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. https://doi.org/10.1093/pnasnexus/pgag008
- 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, CHI '25*, 1–29. https://doi.org/10.1145/3706598.3713171

References (continued)

- Graphite.io (n.d.). *More Articles Are Now Created by AI then Humans. Graphite.io. Retrieved 2 May 2026, from* https://graphite.io/five-percent/more-articles-are-now-created-by-ai-than-humans
- Haider, J., Söderström, K. R., Ekström, B., & Rödl, M. (2024). GPT-fabricated scientific papers on Google Scholar: Key features, spread, and implications for preempting evidence manipulation. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-156
- Helberger, N., De Vreese, C., Bouché, G., Ferrari Braun, A., Drunen, M. V., Kruschinski, S., Mattis, N., Morosoli, S., Naudts, L., Papaevangelou, C., Seipp, T. J., Votta, F., & Weikmann, T. (forthcoming). *Generative AI and Democracy. Study on the advantages and risks of the use of generative artificial intelligence in public debate and democratic processes* (Report for the Council of Europe, Steering Committeee on Democracy (CDDEM) CDDEM(2005)18). AI, Media & Democracy Lab.
- Helberger, N., Miotto, M., & Cools, H. (2025). *Written input for the AI Office Working Group 2 on DIsclosures of deep fakes and certain AI generated texts*. Algosoc and AI, Media & Democracy Lab. https://algosoc.org/uploads/Written-contribution-WG2_1-v01-Sources.pdf
- Hove, M. F., Adler-Nissen, R., Bechmann, A., de Vreese, C. H., Hjorth, F., & Golovchenko, Y. (2024). *Tech-giganter og demokrati: Kunstig intelligens, misinformation og digitale platforme*.
- Jones, R., & Jones, B. (2019). Atomising the News: The (In)Flexibility of Structured Journalism. *Digital Journalism*, *7*(8), 1157–1179. https://doi.org/10.1080/21670811.2019.1609372
- Kang, D., Hashimoto, T.B., Stoica, I., & Sun, Y. (2022). ZK-IMG: Attested Images via Zero-Knowledge Proofs to Fight Disinformation. *ArXiv, abs/2211.04775*.
- Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience*, *24*(11), 103364. https://doi.org/10.1016/j.isci.2021.103364
- Liao, Q., et al. (2025). Impact of artificial intelligence-generated content labels on perceived accuracy, message credibility, and sharing intentions for misinformation. *Journal of Medical Internet Research*. https://pmc.ncbi.nlm.nih.gov/articles/PMC11892328/
- Miotto, M., Priante, A., & Li, T. (2025). Seeing Isn't Believing: How Deepfake Salience Disrupts News Credibility and Sharing on Social Media. *Academy of Management Proceedings*, *2025*(1), 22544. https://doi.org/10.5465/AMPROC.2025.22544abstract
- Perkins, M., Roe, J., Vu, B., Postma, D., Hickerson, D., McGaughran, J., Vietnam, H.Q., & Singapore, J.C. (2024). Simple techniques to bypass GenAI text detectors: implications for inclusive education. *International Journal of Educational Technology in Higher Education, 21 , 53.* *https://doi.org/10.1186/s41239-024-00487-w* .
- Philips, T., & Torres, P. (2024, August 27). Being on camera is no longer sensible": Persecuted Venezuelan journalists turn to A. *The Guardian*. https://www.theguardian.com/world/article/2024/aug/27/venezuela-journalists-nicolas-maduro-artificial-intelligence-media-election
- Piasecki, S., Helberger, N., Morosoli, S., & Naudts, L. (2025). 'I would feel manipulated': Do the transparency provisions in the AI Act give news consumers what they hope for? *Internet Policy Review*.
- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.
- Scharfbillig, M., Lewandowsky, S., Altay, S., van Alstyne, M., Kozyreva, A., Hertwig, R., Lorenz-Spreen, P., DiResta, R., Valenzuela, S., Egidy, S., Quattrociocchi, W., & Orben, A. (2026). *Fractured Reality - How democracy can win the global struggle over the information space,* Publications Office of the European Union, Luxembourg, https://data.europa.eu/doi/10.2760/9358883, JRC144603.
- Schiff, K. J., Schiff, D. S., & Bueno, N. S. (2025). The Liar's Dividend: Can Politicians Claim Misinformation to Evade Accountability? *American Political Science Review*, *119*(1), 71–90. https://doi.org/10.1017/S0003055423001454
- Shadi Samir Chreitehs presentation from eSundhedsobservatoriet: https://2024.e-sundhedsobservatoriet.dk/wp-content/uploads/sites/26/2024/10/oB1-Shadi-Samir-Chreiteh.pdf and https://2024.e-sundhedsobservatoriet.dk/wp-content/uploads/sites/26/2024/06/B1-Shadi-Samir-Chreiteh.pdf
- 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. *New Security Paradigms Workshop*, 17–33. https://doi.org/10.1145/3633500.3633502
- Simchon, A., Edwards, M., & Lewandowsky, S. (2024). The persuasive effects of political microtargeting in the age of generative AI. *PNAS Nexus*, *3*, pgae035. https://doi.org/10.1093/pnasnexus/pgae035.
- Suchman, L. A. (2002). Practice-based design of information systems: Notes from the hyperdeveloped world. *The Information Society*, 18(2), 139-144 https://doi.org/10.1080/01972240290075066
- van Reijmersdal, E., Neijens, P., & Smit, E. (2005). Readers' Reactions to Mixtures of Advertising and Editorial Content in Magazines. *Journal of Current Issues & Research in Advertising*, *27*(2), 39–53. https://doi.org/10.1080/10641734.2005.10505180

CAISA's website

CAISA's LinkedIn