RESPONSE TO THE CONSULTATION ON DRAFT GUIDELINES ON TRANSPARENCY OBLIGATIONS UNDER ARTICLE 50 OF THE AI ACT

June 1, 2026
Authors: Gaston Besanson, Liza Dixon, Natasa Milic-Frayling, Alejandro Saucedo, Gerhard Schimpf
Reviewers: Paolo Giudici, Francisco Medeiros
Editors: Tom Romanoff, Michel Beaudouin-Lafon

1. Executive Summary and Methodological Note

The Association for Computing Machinery (ACM) is the world's longest-established professional society of individuals involved in all aspects of computing. It annually bestows the ACM A.M. Turing Award, often popularly referred to as the "Nobel Prize of Computing." ACM's Europe Technology Policy Committee ("Europe TPC") is charged with and committed to providing sound **technical information** to policymakers and the general public in the service of sound public policymaking. Europe TPC has previously responded to European Union stakeholder consultations in the context of the AI Act[1], the Data Act[2], the Digital Services Act[3,4], the Digital Citizen Principles[5], and the Cyber Resilience Act[6], amongst others[7]. ACM and Europe TPC are non-profit, non-political, and non-lobbying organisations.

Europe TPC supports the European Commission's intent to refine the obligations under Article 50 of the AI Act in order to promote trust, accountability, and integrity in the information ecosystem. This response focuses on those parts of the draft Guidelines where technical feasibility, human-computer interaction, and system architecture materially affect the practical implementation of transparency obligations.

In particular, Europe TPC recommends that the Guidelines:

1. Differentiate compliance expectations by modality, recognising that text, image, audio, and video marking techniques do not currently offer equivalent levels of robustness, interoperability, reliability, and effectiveness.
2. Assess marking and provenance mechanisms across realistic redistribution chains,

[1] https://www.acm.org/binaries/content/assets/public-policy/europe-tpc-comments-ai-consultation.pdf
[2] https://www.acm.org/binaries/content/assets/public-policy/acm-eur-tpc-data-act-comments-13may22a.pdf
[3] https://www.acm.org/binaries/content/assets/public-policy/europetpc-digital-services-act-comments.pdf
[4] https://www.acm.org/binaries/content/assets/public-policy/acm-europe-tpc-dsa-comments.pdf
[5] https://www.acm.org/binaries/content/assets/public-policy/europetpc-comments-digital-principles.pdf
[6] https://www.acm.org/binaries/content/assets/public-policy/acm-europe-tpc-cyber-reslience-comments-pdf
[7] https://www.acm.org/public-policy/public-policy-statements

including editing, screenshotting, compression, re-uploading, and platform-mediated metadata handling.

3. Treat "obviousness" and user notice as dynamic, context-dependent properties affected by habituation, cognitive load, accessibility, and repeated exposure.
4. Clarify that human review of AI-generated public-interest text must be substantive, auditable, and proportionate to scale.
5. Distinguish ordinary disclosed synthetic reproduction, deep fakes within the meaning of the AI Act, and aggravated deceptive conduct such as intentional removal or obscuring of provenance signals.

2. Article 50(2): Technical Feasibility of Marking and Detection

Technical Critique:

- **Delineation by Modality:** In Section 4.2.3 of the draft guidelines ("Compliance with the requirements for technical solution(s): effective, interoperable, robust and reliable"), the Commission explains how technical solutions for marking and detection, which are to be effective, interoperable, robust, and reliable, as required by the AI Act, should be understood (paragraph 74). As the Commission rightly acknowledges in paragraph (78) of this same section, under the current state of the art, no single technique simultaneously meets all four of these requirements. While cryptographic provenance and robust watermarking show promise for high-fidelity spatial and temporal data (images, audio, video), text generation remains highly resistant to robust, imperceptible watermarking.[i] Text manipulation frequently strips metadata, and semantic watermarking is highly susceptible to adversarial evasion[ii] or standard editing practices. ACM ETPC recommends that the guidelines explicitly differentiate compliance expectations based on the structural modality of the data, acknowledging metadata stripping rates that, under current redistribution pipelines, exceed levels compatible with effective downstream detection.
- **Cost, Proportionality, and the Interoperability vs. Robustness Trade-off:** Section 4.2.3, paragraph (79), explicitly allows the cost of implementation as a relevant factor. Several of our recommendations carry non-trivial compliance costs, making proportionality criteria essential. There is an inherent architectural tension between interoperability and robustness. Highly interoperable open standards for metadata are easily stripped (low robustness), whereas highly robust, proprietary watermarking schemes often fail to be interoperable across different detection ecosystems.

Proposed Refinements:

Where the interoperability-robustness trade-off is structurally unavoidable, ACM Europe TPC recommends adding a tiered approach to the state-of-the-art clarifications in paragraph (77):

> *"Where the interoperability-robustness trade-off is structurally unavoidable, a tiered approach shall apply: (i) for high-impact public-interest content (deep fakes of public figures, electoral content, health information, as defined in paragraph (123) on matters of public interest), robustness takes precedence and the marking solution must be designed to survive standard redistribution and editing; (ii) for closed-loop industrial and B2B applications already addressed in paragraph (81), interoperability across enterprise systems is the controlling property; (iii) for general-purpose consumer applications, providers must document the trade-off made and justify it under the proportionality criteria of paragraph (79)."*

3. Articles 50(1) and 50(3): Human-Computer Interaction and the "Obviousness" Exception

Technical Critique:

- **Cognitive Habituation:** For interactive AI systems, Section 3.2.1 of the draft guidelines ("Exception for obvious interaction with an AI system"), paragraphs (39) to (42), proposes an exception to the transparency obligation if the artificial nature of the interaction is obvious to a "reasonably well-informed, observant and circumspect" natural person. From a Human-Computer Interaction (HCI) perspective, obviousness is not static. As Section 3.1.2 ("Information obligation under Article 50(1) AI Act"), paragraph (36), correctly flags regarding "banner blindness," users interacting frequently with agentic systems experience rapid habituation.[iii] What is obvious in an isolated testing environment ceases to be obvious in sustained, evolving interactions.
- **Extension to Article 50(3):** This habituation problem is particularly acute for systems covered under Section 5 of the draft guidelines ("Article 50 (3) AI Act: Emotion recognition systems and biometric categorisation systems"). Deployers in commercial settings (e.g., retail, advertising) who rely on one-time signage will find that such disclosures lose all salience over repeated exposure, rendering the transparency technically present but cognitively invisible. A similar dynamic applies to emotion recognition deployed in transit hubs, public spaces, or healthcare waiting areas, where exposure is involuntary, repeated, and where one-time signage placed at the point of entry rapidly loses cognitive salience over successive exposures. Relying on the subjective assessment of an average user's digital literacy introduces immense compliance uncertainty.

Proposed Refinements:
To establish structural rigour, we propose adding specific objective criteria to paragraph (42) of the draft guidelines:

> *"Obviousness shall be assessed not only by audience perception but also by the presence of objective architectural features that structurally prevent impersonation, including: (i) non mimetic voice cadence or distinguishable synthetic voice markers in voice agents; (ii) persistent, non dismissible user interface elements identifying the AI counterpart throughout the interaction; (iii) machine verifiable AI identity attestations, including those issued via EU Digital Identity Wallets in accordance with Regulation (EU) No 910/2014. Such measures shall be designed and implemented to account for habituation effects, repeated interactions, and realistic conditions of user attention, including high cognitive load and environmental distractions. Obviousness shall therefore be evaluated not only at the point of initial interaction, but with regard to the continued effectiveness of transparency mechanisms throughout the duration and repetition of use"*

4. Article 50(4): Deep Fakes, Text Generation, and Editorial Responsibility

Technical Critique:

- **Ontological Distinction of Deep Fakes and Digital Sovereignty:** In Section 6.1.1 of the draft guidelines ("The notion of 'deep fake'"), the current text risks conflating benign AI-generated reproductions with adversarial deepfakes. A structural distinction must be established between a disclosed synthetic reproduction, such as authorised digital twinning for preservation or accessibility, and a "deep fake." Mandating transparency for all synthetic reproductions is technically sound and supports digital biometric integrity, the digital extension of individual self-ownership. However, the regulatory definition of a deepfake should be strictly reserved for instances of intentional adversarial unmarking, where provenance markers have been deliberately circumvented, obscured, or removed for the purposes of deceit.
- **Scaling and "Neutral" Inference:** In Section 6.2.3 ("Exception from the transparency obligation for text under human review or editorial control and editorial responsibility"), paragraphs (126) and (127) of the draft guidelines provide an exception for AI-generated text published on matters of public interest if it undergoes human review, correctly ruling out "superficial, solely formal or procedural checks." However, these definitions must account for the reality of automated scaling. If an enterprise architecture deploys AI to

generate thousands of variations of public-interest text (e.g., hyper-localised public safety warnings), a human reviewer may physically approve the batch without substantive engagement per item.

- **Definition of Standard Editing and Substantial Alteration:** In Section 4.3 of the draft guidelines ("Exceptions to the obligations under Article 50(2) AI Act"), paragraphs (84), (85), and (86) exempt standard editing that improves readability or format without altering semantics. Given the deployment of advanced LLMs, restructuring a sentence for readability can inadvertently shift its neutral inference or factual weight, and standard editing operations frequently degrade watermark integrity.[iv] Paragraph (86) gives examples but lacks a testable, operational definition of "substantial alteration."

Proposed Refinements:

To prevent the automation of misinformation under the guise of blanket editorial responsibility, we propose inserting the following limitation into paragraph (127)of the draft guidelines :

> *"Editorial control over AI-generated text at scale shall not be presumed substantive where the volume of items reviewed per human reviewer exceeds thresholds established by sectoral codes of practice under Article 50(7), or where the median time of substantive engagement per item falls below what is necessary for fact-checking and source verification."*

To ensure semantic drift is properly categorised and disclosed, we propose adding the following operational definition to paragraph (85) of the draft guidelines:

> *"For the purposes of paragraph (85), an alteration should be considered substantial where it changes the entailment relationship between the input and output content under standard natural-language inference evaluation, or where it adds, removes, or replaces semantically load-bearing elements (named entities, factual claims, quantitative values, depicted persons or objects)."*

5. Operationalisation through the Code of Practice (Article 50(7))

ACM Europe TPC observes that several of the technical specifications proposed above, including operational thresholds for substantive editorial review, taxonomies of semantic alteration, and benchmarks for the four quality requirements under Article 50(2), are appropriately operationalised through the Code of Practice mechanism foreseen in Article 50(7) and described in Section 8.1of the draft guidelines ("Effects of adhering to a code of practice

assessed as adequate"), paragraphs (135) to (138). Attempting to hardcode dynamic technical thresholds directly into static guidelines risks rapid obsolescence. The Code of Practice provides the necessary agile regulatory coverage to adapt to the evolving capabilities of frontier systems. In addition, objective and structurally verifiable criteria of the kind proposed above materially reduce enforcement ambiguity for the market surveillance authorities responsible for Article 50 under Section 8.2 ("Market Surveillance Authorities") and Section 8.3 ("Penalties"), paragraphs (139) to (140).

6. Conclusion

While the core transparency obligations are established by the AI Act itself, these guidelines will dictate how market surveillance authorities assess and enforce compliance. To ensure this enforcement remains operationally viable and technically feasible for global data architectures, the Commission must anchor its regulatory expectations to empirical detection limits, acknowledge modality-specific trade-offs, and rely on structurally verifiable interaction criteria rather than subjective assessments.

References

[i] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A Watermark for Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning (ICML '23)*, Honolulu, Hawaii. PMLR, 17061–17084

[ii] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can AI-Generated Text be Reliably Detected? *Transactions on Machine Learning Research* (2023). https://arxiv.org/abs/2303.11156

[iii] Nielsen Norman Group. 2018. Banner Blindness Revisited: Users Dodge Ads on Mobile and Desktop. Retrieved June 1, 2026 from https://www.nngroup.com/articles/banner-blindness-old-and-new-findings/

[iv] [4] Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. 2023. Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models. arXiv:2311.04378. https://arxiv.org/abs/2311.04378