WITNESS
SEE IT FILM IT CHANGE IT

WITNESS Comments regarding the EU AI Act's First Draft Code of Practice on Transparency of AI-Generated Content

WITNESS welcomes the opportunity to provide input on the First Draft Code of Practice on Transparency of AI-Generated Content as members of the Working groups 1, on Rules for marking and detection of AI-generated and manipulated content applicable to providers of AI systems (Article 50(2) and (5) AI Act), and Working Group 2, on Rules for labelling of deepfakes and AI-generated and manipulated text applicable to deployers of AI systems (Article 50(4) and (5) AI Act).

Section 1: Rules for marking and detection of AI-generated and manipulated content applicable to providers of AI systems (Article 50(2) and (5) AI Act)

Commitment 1: Multi-layered Marking of AI-Generated Content

Measure 1.1: Machine-readable marking techniques
Sub-measure 1.1.1: Marking techniques for content that permits metadata embedding
Sub-measure 1.1.2: Marking techniques interwoven within the content
Sub-measure 1.1.3: Fingerprinting or logging facilities (where necessary)

- We welcome the multi-layered marking requirement as it is needed to facilitate transparency that can be effectively interpreted by real-world users, as required by Article 50.
- The **multi-layered marking requirement is both feasible and appropriate** in its current version. The text could encourage the use of emerging techniques without weakening the multi-layered obligation, and also acknowledging the state of the art's ability to meet regulatory requirements (as it currently does).
- Privacy considerations need to be incorporated into this Code; at a minimum:
    - Personally identifiable information should not be embedded in markings or provenance data by default;
    - Where any user-related or contextual data is strictly necessary, it must be data-minimised, protected, and aligned with existing data protection law, including the GDPR;
    - Control over such data should rest with the appropriate data controllers and rights-holders, not be broadly exposed or centralised.
- To avoid role confusion, the draft could be clearer that the obligations under this measure apply to providers of AI systems, not models. **HOWEVER**, the underlying requirements should be upheld as **enabling conditions** for downstream compliance by deployers. Ie: often, marking techniques (e.g., watermarking) work best at the model or training stage, and by framing these as "enabling conditions," it is clarifying that

upstream design features allow downstream deployers to meet disclosure obligations without creating separate legal duties for the model. **This clarification should enable effective transparency without shifting the legal responsibility to model providers**. This is contemplated in Article 50 of the EU AI Act:

- Article 50(1) specifies that "Providers shall ensure that AI systems intended to interact directly with natural persons are ***designed and developed*** in such a way that the natural persons concerned are informed that they are interacting with an AI system"
- Recital 133, and Recital 136 emphasize that provider measures **facilitate downstream deployer transparency and effective implementation**, ensuring disclosures are technically feasible, enforceable, and meaningful from a rights perspective. Without multi-layered marking by design, deployer disclosure is often infeasible, undermining transparency.

Measure 1.2: Marking techniques for specific modalities

Sub-measure 1.2.1: Provenance certificate for AI-generated text and other content that does not allow secure embedding of metadata

Sub-measure 1.2.2: Marking of multimodal content

- We applaud section 1.2.2 as this specification for multimodal content better reflects how content is actually made and edited in the real world. In particular, ensuring that markings be recognisable even when only one or a subset of modalities have been altered or exchanged is necessary for downstream transparency.

Measure 1.3: Structural Marking for open-weight AI models and systems

Cf.1.1

- To avoid role confusion, the draft could be clearer that the obligations under this measure apply to providers of AI systems, not models. **HOWEVER**, the underlying requirements should be upheld as **enabling conditions** for downstream compliance by deployers.
    - CRITICALLY, for open-weight AI models, providers can only be fully compliant by implementing markings at training. Once the weights are released, anyone can generate content outside the law's scope, undermining transparency and the intent of Article 50. Marking at training is therefore the critical point for effective enforcement and protection of the information ecosystem.

Measure 1.4: Marking techniques at the level of the generative AI model

- Although AI models are out of scope, this provision is necessary as a prerequisite for effective downstream transparency (for example, watermaking at the point of training). The draft should clarify that it applies to AI system providers, not model providers, noting that marking at the point of training is an **enabling condition** for downstream compliance.

- We welcome the call for SMEs or SMCs that are providers of AI systems to use one or more generative AI models which already mark the outputs in a manner compliant with the relevant measures in Section 1 of the Code, while also clarifying that they still retain responsibility to ensure that all AI-generated or manipulated outputs are suitably and compliantly marked.

Measure 1.5: Non-removal of machine-readable marking

- We welcome this provision and reject claims that it is out of scope or unduly burdensome. It does not impose obligations beyond the AI system itself. Specifically, provision 1.5(b) does not require AI systems to guarantee that markings cannot be removed; rather, it addresses terms, policies, conditions, and documentation that help mitigate the risk of removal.

Measure 1.6: Transparency of the provenance chain

- We welcome this provision as it facilitates compliance with the regulation and its intent. In particular, because it leans into the notion of the 'recipe' of the content that more accurately reflects how content is actually made and edited in the real world (as opposed to an AI/not AI binary).
- AI is part of a process of creation over time, including multiple AI and human ingredients combined, mixed in no particular order and in no specific stages of the content's lifespan. For the Code to facilitate compliance with regulation, it should therefore leverage a system that can reflect this reality; that is, multi-layered marking that includes the provenance of a content.

Measure 1.7: Functionality for perceptible markings (for deep fakes and other content)

- While the draft could be clarified to explicitly reference AI systems, the underlying requirements are appropriate and within scope.
- Requiring functionalities to include perceptible markings supports downstream compliance and reflects real-world literacy needs (ie: users are more likely to correctly interpret marks from major AI system providers than from a diverse set of deployers).
- This is also an issue of **enabling conditions** that facilitate downstream compliance. For example, the leading initiative for metadata provenance is the C2PA, which has a gated ecosystem (via its Conformance Program) that may be less accessible to a range of Deployers than to providers of AI systems. By including these requirements upstream, providers help ensure deployers can effectively comply.

Commitment 2: Detection of the Marking of AI-Generated Content

Measure 2.1: Enable detection by users and other third parties

- We welcome this measure and find that it is appropriate and within the scope of the regulation.

Measure 2.2: Detectors for already marked AI-generated content produced by a generative AI model

- Although AI models themselves are out of scope, the intent of this draft remains unchanged if the text simply refers to providers of AI systems. This can be achieved by removing the phrase "[…]who are also providers of generative AI models[…]".

Measure 2.3 Forensic detection mechanisms

- Although AI models themselves are out of scope, this measure could be redrafted to refer to AI systems and clarify that the systems they provide include forensic detection functionalities that do not rely on active AI markings.
- Upstream implementation of forensic detection functionalities makes this more enforceable and it facilitates user literacy and interpretability. For example, model developers are exceptionally placed to create post-hoc detection systems (eg. via model attribution techniques).
- This does not preclude deployers from also being responsible for ensuring forensic detection in their outputs.
- Added to that, we would like to reinforce the need for collaboration amongst different groups of stakeholders. Upstream implementation is useful and practical but we should not ignore the responsibility that lies in groups such as downstream developers.
- The intent of the regulation is to ensure that natural persons are made aware of their interactions with an AI system, and this requires a holistic approach to detection that considers reception and socio-technical dimensions at every stage of content creation and sharing. WITNESS developed the TRIED benchmark to specify what those elements that can ensure compliance with the regulation are.

Measure 2.4: Human-understandable and accessible disclosure of verification and detection results

- This measure is appropriate and within scope for providers of AI systems, as it concerns system-level capabilities. However, there may be use cases where mandating this measure would not be proportionate, such as industrial or other AI systems operating in closed or purely internal environments. Rather than excluding the provision, the Code could clarify its applicability by explicitly carving out such cases where no interaction with the public or external recipients occurs.
- There is a need for concrete, shared standards when determining what constitutes human-understandable and accessible disclosure, rather than leaving such judgments solely to the discretion of signatories. Without clearer benchmarks, disclosures risk becoming formally compliant yet substantively opaque to the very audiences they are meant to inform. Meaningful accessibility cannot be defined in the abstract; it depends on the capacities, contexts, and needs of affected users and stakeholders. Accordingly, processes for defining and evaluating disclosure should include structured engagement with users, civil society, and other relevant stakeholders, enabling these standards to be co-developed rather than unilaterally set. Such participatory approaches would help

ensure that disclosure practices are not only technically accurate, but genuinely comprehensible and useful in practice.

Measure 2.5: Support literacy for AI content provenance and verification

- Appropriate and within scope for providers of AI systems.

Commitment 4: Testing, verification and compliance
Measure 4.1: Compliance framework

- Appropriate and within scope for providers of AI systems.

Measure 4.2: Testing, verification and monitoring

- Appropriate and within scope for providers of AI systems. Testing, verification and monitoring are necessary components to ensure and facilitate compliance and enforceability.

Working group 1: Additional questions

Shared verifiers for AI-generated content
How could a shared verifier for AI-generated content (defined as a detector or verifier for markings originating from multiple providers of AI systems or models) be implemented technologically, considering economic and security constraints? Should such verifiers be centralised or decentralised? (650 character(s) maximum)

There could be different types of shared verifiers:

1. A validator of structured, consistent marking technique. For example a C2PA validator.
2. A company or model-based detector.
3. A post-hoc AI detector

The first two can become shared verifiers in a straightforward fashion. Company or model-based detectors would just need to implement security measures so that it is not undermined (although there are security considerations to bear in mind, these are not an argument to limit accessibility, as it undermines the intent of the regulation).

For these two types of shared verifiers, we strongly encourage a decentralized system (implied with C2PA). This is necessary to facilitate detection and comply with the law, especially within workflows. Centralized detection would not enable compliance with the law as it places the burden on the user to leave a deployer's interface.

Post-hoc detection may be out of scope of this regulation, though deployers could be encouraged to work with post-hoc detectors that follow a TRIED benchmark.

Other technical recommendations

Are there any other technical considerations that you would like to raise and recommendations for public infrastructure and services? (650 character(s) maximum)

To protect privacy, the Code should clearly state that:

- Personally identifiable information should not be embedded in markings or provenance data by default;
- Where any user-related or contextual data is strictly necessary, it must be data-minimised, protected, and aligned with existing data protection law, including the GDPR;
- Control over such data should rest with the appropriate data controllers and rights-holders, not be broadly exposed or centralised.

Section 2: Rules for labelling of deepfakes and AI-generated and manipulated text applicable to deployers of AI systems (Article 50(4) and (5) AI Act)

Part A: General Commitments

Commitment 1: Disclosure of Origin of AI-Generated and Manipulated Content based on a Common Taxonomy and an Icon

Measure 1.1: Implement a common taxonomy

- With regards to the common taxonomy, we would welcome the adoption of less subjective terms referring to the content that is targeted by the COP. As highlighted in our previous submissions, the final text must refrain from enabling perceptions of content such as the "made by AI" or "made by human" binary, and avoid the misconception that all AI-manipulated or generated content is done solely for deceptive purposes. Lastly, AI-generated and manipulated media are used across a wide range of contexts, and span many formats, including images, video, audio, avatars, and multimodal combinations, and it would be important for the COP to acknowledge that also through the common taxonomy.
- When deploying media literacy efforts, the taxonomy or lack thereof should not be used to undermine or bolster content.

Measure 1.2: Applying a common icon for AI-generated and manipulated content

Sub-measure 1.2.1: Pending development of an EU-wide icon

- We welcome the idea of a common icon as long as it comes with more information about the content the user is being exposed to. Added to that, the icon should be interactive, recipe-based and take into account different types of media and distribution methods. Lastly, given technical considerations towards the application to types of content such as audiovisual.
- This requires strong interoperability and a commitment to preserve this information that applies to providers, deployers and potentially other actors in order to be effective.

Providing structured, persistent, and interoperable *recipe information (more comprehensive history of its provenance) is essential for* building public trust, supporting accountability, and enabling users to interpret AI-mediated content accurately across contexts. Understanding the recipe for AI and human contribution in what we consume is, at its core, about knowing the ingredients of AI and human-content, and confirming how they were added in. And sometimes, it's about knowing the chef.

Sub-measure 1.2.2: EU common icon

- The standardized approach to the icon is a positive development, as it can foster greater interoperability, enable consistent implementation, and prevent the emergence of fragmented ecosystems with multiple icons or labeling practices. The EU common icon should be interactive, recipe-based, and designed to accommodate different types of media and distribution channels.

- In this context, we would like to reiterate the importance of providing information related to the "recipe." Understanding the balance between AI and human contribution in the content we consume is fundamentally about knowing the respective "ingredients" and how they were incorporated. We also wish to emphasize the importance of a layered information system to ensure that transparency remains usable, proportionate, and effective across diverse environments.

- Regarding the use of gradients in the AI icon, the absence of a clear and effective taxonomy risks placing an unnecessary burden on end users, who may struggle to understand what each step in the gradient represents (AI-generated vs. AI assisted and the translations). In contrast, the adoption of a single, standardized, simpler interactive icon (similar to an information icon) could prove more effective by facilitating comprehension for end users, supporting the broader ecosystem, and strengthening media literacy efforts.

- We agree with the principle of interactivity and see it as a meaningful effort to provide users with additional context, which can be layered to offer more detailed information as needed. However, a label that appears clear and unmistakable in one context may become far less visible or intuitive when the same content is shared on short-form video platforms, messaging services, or platforms with different design conventions, or when content is edited, clipped, or repurposed. Moreover, the perceived clarity of a label or disclosure is influenced by factors such as age, language, cultural expectations, and varying levels of media literacy, as well as by resharing and remixing practices.

Commitment 2: Compliance, training and monitoring
Measure 2.1: Internal compliance

- It is crucial that the COP anticipates and addresses scenarios in which the icon is removed or not included on the basis of having no detectable markings related to the

common taxonomy. SORA is a recent example where watermarks were easily stripped using dedicated tools within days of release. The urgency of this issue is underscored by the widespread availability of AI video tools from companies such as OpenAI and Google, where content is rapidly created, often detached from its original context, and can shift from satire, entertainment, or communication to malicious or deceptive use.

Commitment 3: Ensure Accessible Disclosure for all Natural Persons

Measure 3.1: Accessibility of the labelling of deepfakes and AI-Generated or manipulated text

- Provenance and Detection within platforms: The encouragement to make detection mechanisms directly available in distribution and communication platforms is the right direction, though the language here is still too lenient in terms of expectations.
- Still on this issue, we would like to highlight the relevance of the **multilayered approach** as it **can help ensure that the disclosures persist even when content is reshared, edited, or moved across platforms, without relying solely on visible labels**. This means that visible labels can provide simple, lightweight cues, while cryptographically signed provenance metadata—such as C2PA—anchors authenticity in a tamper-evident, machine-readable layer. Even when surface labels are altered or removed, embedded signatures allow platforms and tools to verify provenance. By combining the steps above, we believe that the disclosures model adopted by deployers and developers can make this process more meaningful, accessible, and resilient—while also addressing the higher levels of content available due to the AI Slop
- Such an approach gives users the right level of information at the right time, while enabling deployers and developers to move beyond simplistic distinctions between "benign" and "malicious" content. It acknowledges the diversity of AI-generated and AI-modified expressions. The goal is not to police intent, but to provide consistent, trustworthy signalling at scale so people can understand how content was created and assess it appropriately.

- Lastly, the COP could also take into account some of the learnings emerging through the C2PA User Experience Guidance for Implementers, a set of UX recommendations aimed to define best practices for presenting C2PA provenance to consumers in order to empower them to understand where it came from and decide how much to trust it.

Part B: Specific Commitment and Measures for Deepfakes

Commitment 4: Specific Measures for Deepfake Disclosure

Measure 4.1: Internal processes for consistent classification of Deepfake Content

- There is a need for concrete, shared standards when determining what constitutes human-understandable and accessible disclosure, rather than leaving such judgments solely to the discretion of signatories. Without clearer benchmarks, disclosures risk becoming formally compliant yet substantively opaque to the very audiences they are meant to inform. Meaningful accessibility cannot be defined in the abstract; it depends on the capacities, contexts, and needs of affected users and stakeholders. Accordingly,

processes for defining and evaluating disclosure should include structured engagement with users, civil society, and other relevant stakeholders, enabling these standards to be co-developed rather than unilaterally set. Such participatory approaches would help ensure that disclosure practices are not only technically accurate, but genuinely comprehensible and useful in practice.

Measure 4.3: Apply Appropriate Disclosure for Creative Works

- *WITNESS previous work - such as the report launched in 2021, named "**Just Joking: Deepfakes, Satire, and the politics of Synthetic media**"; and an article published in 2023 - help highlight the importance of dealing with labels as an inherent part of the content. Added to that, we also recommend that the COP deals with this issue as more than an add-on functionality and set a minimum requirement for disclosures. Creative work should also not be exempted from the disclosures obligation, especially if the common EU Icon is more generic (i.e. the i for information), as this can enable compliance and enforceability that may not be achievable if exceptions for subjective understandings of satire and art are made.*
- *A label that feels clear and unmistakable in one setting may become far less noticeable or intuitive when the same content moves to short-form video apps, messaging channels, or platforms with different design norms. Added to that, the perceived obviousness of a content and/or disclosure can also be shaped by factors such as age, language, cultural expectations, and varying levels of media literacy of the target audience, as well as potential reshares and remixes of existing content. This emphasizes the need for multi-layered marking techniques to help protect satire/creative content on the visible or audible layer, and enable underlying disclosure. It is worth highlighting here again the need to ensure that these techniques, while required, do not infringe on privacy, but rather focus on non-personal provenance.*

****

Relevant Materials from WITNESS

- Regulating Transparency in Audiovisual Generative AI: How Legislators Can Center Human Rights summarizes our key understandings of the transparency and labeling landscape
- C2PA Harms Assessment: WITNESS co-chairs the Threats and Harms groups and identifies/proposes solutions for potential harms, including to fundamental rights, such as privacy, access limitations, or potential weaponization of approaches by malicious actors including governments
    - https://spec.c2pa.org/specifications/specifications/2.0/security/Harms_Modelling.html
    - https://spec.c2pa.org/specifications/specifications/2.0/security/_attachments/Due_Diligence_Actions.pdf
- Embedding Human Rights in Technical Standards: Insights from WITNESS's Participation in the C2PA, identifies key learnings about ensuring fundamental rights in the type of output labelling and disclosure frameworks as identified in Article 50

- Just Joking: an ongoing project focused on understanding satirical deepfake content, a usage identified in Art 50.4
- The Thorny Art of Deepfake Labeling reviews lessons learned from creative communities and satire on context-driven AI labeling
- The TRIED Benchmark is focused on ensuring complementary frameworks for assessing quality of post-hoc detection, used when content does not include relevant markers or previous disclosure of synthetic output, but disclosure needs to be added after the fact.

Additional resources on the work we've done since 2018 on deepfakes and synthetic content, as well as labelling and disclosure, are gathered at gen-ai.witness.org.