The Problems with "General Purpose AI Detectability"

On the EU's new "Code of Practice on Transparency of AI-Generated content"

Paul Friedl

08.07.2026 doi: 10.59704/bbba65823b833130

As AI-generated media flood our information ecosystems, detecting synthetic content has become an urgent regulatory challenge – in fact, not one challenge but many, as synthetic media breeds problems across a range of digital contexts, including deepfakes and disinformation, scamming, and content moderation. The EU's new "Code of Practice on Transparency of AI-Generated Content" – the first concrete articulation of Article 50(2) AI Act, the EU's approach to AI-content detection – gives sensible answers to several open questions and will advance the global regulatory debate. Unfortunately, however, it cannot escape the structural shortcoming at the core of Article 50(2): the premise of "general-purpose AI detectability," which treats detection as a "fixable" technical problem rather than a context-dependent, political one demanding iterative, multi-actor coordination. That diagnosis points towards a different response: rethinking synthetic-media regulation, and kindred problems of internet governance, around institutionalized, regulator-led forums that oblige stakeholders to keep communicating and collaborating.

Synthetic Content Transparency: Article 50 AI Act and the EU's new Code of Practice

Establishing AI detectability is no simple problem: it has to deal with a quickly evolving technological landscape and a range of diverse actors – AI developers, platforms, researchers and more – whose interests only partly overlap. Fortunately, regulators have nonetheless begun to act, adopting legislation to help increase synthetic media detection capabilities. One of these regulators, alongside China and India, is the EU, which, in Article 50(2) AI Act, obliges "[p]roviders of AI systems, including general-purpose AI systems, generating synthetic audio, image, video or text content, [to] ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated". Furthermore, "[p]roviders shall ensure their technical solutions are effective, interoperable, robust and reliable as far as this is technically feasible".

With the adoption of the "Code of Practice on Transparency of AI-Generated Content" on 10 June 2026 – a guidebook developed by appointed, yet independent academic experts together with over 100 stakeholders from industry, academia and civil society – these obligations have now found a concrete articulation. The Code of Practice lays down a framework of commitments and measures – some of which obligatory, some optional – that can prove adherence to Article 50(2). If the Commission approves the Code, which seems likely, adherence to the Code will create a presumption of compliance with Article 50(2). Although providers are also free to adopt their own, alternative compliance strategies, the legal certainty provided by this presumption will in all probability confer high practical relevance on the Code.

As the remainder of this contribution argues, the Code, for all it gets right, exhibits a number of questionable blindspots and, perhaps inevitably, fails to overcome the structural shortcoming at the core of Article 50(2).

What the Code gets right

One aspect that will be central to the success of any AI detectability strategy is the issue of technical implementation. Synthetic media can only be detected if appropriate technical tools are available to successfully identify AI generations. The issue is a complex one: it is unlikely that concerned actors will ever possess technologies that can ensure a robust marking

of all covered forms of synthetic media; as of today, they don't. The Code obliges AI providers to implement at least two layers of machine-readable marking, focusing on the currently dominant approaches of metadata addition and imperceptible watermarking. However, it also calls on providers to continually test and overhaul their tools to at least attempt to maintain effectiveness and robustness vis-a-vis evolving adversarial challenges – a generally convincing approach.

The Code also takes several welcome steps to advance the coordination between AI developers and other actors involved in content propagation and consumption, which is pivotal to the (inherently collaborative) project of synthetic media detectability. These include the duty for AI providers to prohibit users, through their acceptable use policies, from intentionally removing or tampering with markings and encouraging them to "collaborate with relevant actors within the ecosystem to make their detection solution directly available within distribution and communication platforms". While this does not reach the level of integration implemented by China's regulation on AI content transparency, which extends marking and labelling duties to providers of generative AI systems as well as platform intermediaries and end users, it still marks a significant, if tentative, step toward the cross-actor coordination on which synthetic media detectability depends.

The blindspots

Unfortunately, the Code also suffers from a number of blindspots. First of all, it insufficiently considers that any successful AI detection strategy needs to have special consideration for the problem of malicious actors. Whereas AI slop and other "innocent" synthetic media propagation have risks, the dangers posed by malicious actors willing and able to strategically bypass detection mechanisms (e.g. for disinformation) are not only more grave, but also harder to mitigate. Synthetic media identification, therefore, is structured by what has adequately been called a "detection dilemma": the more broadly accessible a detection technology becomes, the more easily it can be circumvented. This dynamic is compounded by an asymmetry of cost and effort: producing convincing synthetic media has already become incredibly cheap and fast, while marking and detecting it remains technically demanding and expensive. While the Code is not blind to this dynamic and the related insight of an "AI detection arms race", its default is broad public availability: detection solutions are to be made available to the general public, free of charge, with encouragement toward open specifications. Staggered access, i.e. restricting detection to (verified) high-stakes users while accounting for security and equity concerns, is reserved for the case of free-form text watermarking. The result is a posture of "default public access" for the high-fidelity modalities (photo, video, audio) where determined bad actors stand to gain the most.

What is more, the Code prematurely discounts the value of passive, forensic AI detection tools, by treating them as a purely optional measure rather than a potential compliance necessity.

The Code is right in highlighting the current unreliability of forensic AI detection tools and their inferiority to active marking technologies. However, especially against malicious actors, which can strip synthetic media of embedded detection signals or make use of systems that do not apply them in the first place, forensic detection tools can be the only remedy. Forensic detection arguably fits poorly within the AI Act's provider-centric design: developing synthetic content identifiers, in particular model-agnostic ones, is a duty one could just as readily locate with content propagators, as is the case under India's new intermediary-centric regulation on AI detection. However, as actors with highly privileged insight into the characteristics and telltales of AI generations, AI systems providers, especially large ones, are well-equipped to share in the responsibility to build effective forensic detection tools.

A deeper misconstruction - and a way forward

These shortcomings ultimately reflect a deeper misconstruction at the core of the AI Act's approach to synthetic media detectability. Article 50(2) prescribes a horizontal, uniform, one-sided, risk-agnostic, techni-

cal "fix" to what in reality are diverse issues of digital deception, the responses to which are necessarily context-dependent and "political" and require collaboration and coordination between a number of different actors, including AI systems providers, intermediaries, public authorities, researchers and NGOs. The Code even attempts to inspire such networking – encouraging AI providers to "foster collaboration" with deployers and users of their systems – but of course is formally reined in by the silo of "AI regulation". Whether these synapses will be established by other regulations, e.g. the DSA's obligations on platform and search engine providers, is an open question, but seems dubious. Unfortunately, the structure of Article 50(2) and its too broad, and therefore too narrow, perspective of 'general-purpose AI detectability' seem to have prevented the Code of Practice from engaging with the work that actually matters: sensibly delineating which forms of AI-generated content are genuinely deceptive, and devising the nuanced, socio-technical responses capable of countering them.

These insights might help inform future legislation. They also point to a potentially more effective and sustainable method for addressing multi-actor digital governance issues, namely one that recognizes that many such issues, including synthetic media harms, cannot be discharged once and for all by a discrete techno-regulatory intervention, but rather require continuous and evolving multi-party collaboration. Such a framework could be set up through an institutionalized, regulator-led forum that obliges the relevant actors to communicate and collaborate, and in which decisions regarding e.g. the development of and access to detection technologies are regularly revisited in light of evolving real-world problems and priorities. The Code makes a first, rudimentary step in that direction, by announcing a "taskforce … to facilitate … collaboration between Signatories, other actors in the value chain and relevant stakeholders through regular meetings", which Signatories are "encouraged" to participate in. Yet, if the preceding analysis is correct, such a forum belongs at the core of synthetic media regulation, not in an optional taskforce.

ABOUT THE AUTHOR

Paul Friedl

Dr. Paul Friedl is a legal researcher with a focus on data and technology law. He is currently a postdoctoral researcher at Humboldt-University Berlin.

SUGGESTED CITATION

Friedl, Paul: The Problems with "General Purpose AI Detectability": On the EU's new "Code of Practice on Transparency of AI-Generated content", Verfassungsblog, 08.07.2026, doi: 10.59704/bbba65823b833130.

Licensed under CC BY-SA 4.0.

Langzeitarchiviert bei IntRechtDok unter: doi.org/10.17176/20260708-190510-0.