KIRKLAND & ELLIS
Kirkland Alert

Illuminating AI: The EU’s First Draft Code of Practice on Transparency for AI-Generated Content

17 February 2026

On 17 December 2025, the European Commission published a First Draft Code of Practice on Transparency of AI-Generated Content. The draft code is designed to aid compliance with the obligations in Article 50 of the EU AI Act for: (i) providers to mark AI-generated or manipulated content in a machine-readable format and (ii) users who deploy generative AI systems for professional purposes to clearly label deepfakes and AI-text publications on matters of public interest.

Background

The transparency requirements set out in Article 50 of the EU AI Act are designed to make it clear when content has been generated or altered by AI, including so-called deepfakes. According to Article 3(60) of the act, a “deepfake” refers to any image, audio or video content created or modified by AI that imitates real people, objects, places, entities or events in a way that could mislead someone into believing it is genuine.

Article 50(2) of the act requires providers of generative AI systems to ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated. The employed technical solutions must be effective, interoperable, robust and reliable as well as 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.

Article 50(4) of the act requires deployers of AI systems that generate or manipulate image, audio or video content **constituting a deepfake**, to disclose that the content has been artificially generated or manipulated, except where the use is authorised by law to detect, prevent, investigate or prosecute a criminal offence. Where the content forms part of an evidently artistic, creative, satirical, fictional, or analogous work or programme, the disclosure obligation is limited to an appropriate manner that does not hamper the display or enjoyment of the work.

Article 50(4) of the act also requires deployers of AI systems that generate or manipulate text which is **published with the purpose of informing the public on matters of public interest** to disclose that the text has been artificially generated or manipulated, except where the use is authorised by law to detect, prevent, investigate or prosecute a criminal offence. This obligation also does not apply 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.

Article 50(5) of the act requires information disclosed for the purposes of the above to be provided to the individuals concerned in a clear and distinguishable manner, at the time of the first interaction or exposure (if not earlier).

Draft Code of Practice

The purpose of the draft code of practice is to serve as a guiding document for demonstrating compliance with these obligations and to enable the competent market surveillance authorities to assess compliance of providers of generative AI systems who choose to rely on the code to demonstrate their compliance. However, whilst adherence to the code is a step towards compliance, it is not to be treated as conclusive evidence of compliance with the act.

The draft code consists of two sections. The first section covers the marking and detection rules applicable to providers of generative AI systems, while the second section covers the rules that apply to deployers as regards labelling deepfakes and AI-generated/manipulated text on matters of public interest.

The code sets out a number of commitments and measures that signatories to the code will be expected to implement in order to comply with the marking, detection and labelling requirements.

Section 1: Marking and detection rules applicable to providers of generative AI systems

- **Commitment 1: Multilayered marking of AI-generated content** — To fulfil their obligations under Article 50(2) of the act to mark in a machine-readable manner the outputs of generative AI systems, signatories will commit to implement a multilayered approach of active marking techniques which can be implemented at different stages of the value chain (e.g., model providers) and can also be provided by third parties (e.g., providers specialised in transparency marking techniques).

  The code explains marking techniques that can be used to guarantee provenance depending on whether or not the content allows secure embedding of metadata and refers to the need to ensure that marking is retained and not altered or removed. To ensure the transparency of the provenance chain, signatories will be expected to record and embed through content marking the origin and provenance chain from AI-assisted or (partially) modified content to fully AI-generated content where technologically possible.

  Providers who sign up to the code will also be expected to facilitate deployers' compliance with the labelling requirements for deepfakes and other content by providing systems which allow deployers to directly — upon generation of the output — include a perceptible mark or label in the content enabled by default. Signatories will also implement supporting measures for display of labels and provenance metadata that enable deployers, platforms and websites to implement display practices and policies that are appropriate for their use cases.
- **Commitment 2: Detection of the marking of AI-generated content** — To fulfil their obligation under Article 50(2) of the act to ensure that the outputs of their AI system(s) are detectable as AI-generated or manipulated, signatories will commit to implementing a number of measures to enable the detection of text, image, video or audio content, or a combination thereof, as generated or manipulated by their AI system or model. These include providing an interface (e.g., API or user interface) free of charge, or a publicly available detector to enable users and other interested parties to verify (with confidence scores) whether content has been generated or manipulated by their AI system or model. Further, in the event of a signatory going out of business, they will be required to make detectors available to the relevant market surveillance authorities to ensure for the detection of legacy content generated or manipulated by their AI system or model.

  To facilitate compliance by downstream providers, signatories will provide detection mechanisms for the content generated or manipulated by their models prior to the model's placement on the market. They will also commit to collaborate with competent market surveillance authorities and other relevant parties on the detectability of outputs; embed in the results of their marking and detection solution human-understandable explanations; and provide documentation, training materials, and other relevant information to support deployers and other users in making informed decisions on what marking and verification tools they may use. In addition, signatories are encouraged to collaborate with organisations (particularly academia and civil society organisations) to foster greater understanding and awareness around AI content provenance and verification.

- **Commitment 3: Measures to meet the requirements for marking and detection techniques** — Signatories will implement marking and detection solutions that are computationally efficient and low-cost, that ensure real-time application and that are capable of preserving the quality of the generated content. Marking and detection solutions should be reliable and aligned to the state of the art; and signatories should be able to demonstrate low false-positives and false-negatives on samples of AI-generated and human-authored content unseen during the training and development of their AI models or systems.

  Such solutions should also achieve a **high level of robustness** of the marking technique to common alterations (typical processing operations such as mirroring, cropping, compression, screen capturing, paraphrasing, character deletions, changes in image or video resolution, pitch shifting, time stretching, or change of format) and adversarial attacks (such as copying, removal, regeneration, modification and amortisation attacks on the markings). In relation to adversarial attacks, signatories will be required to assess the robustness of their security and further ensure standard security practices are applied to their marking and detection mechanisms to prevent and counteract potential attacks.

  Signatories will also ensure that their marking and detection solutions are **interoperable** across distribution channels and technological environments, regardless of the application domain or context. In this regard, signatories, including SMEs and SMCs, are encouraged to make use of relevant content-marking standards that emerge from international and European standardisation organisations, alongside widely adopted technical standards.
- **Commitment 4: Testing, verification and compliance** — To effectively fulfil and demonstrate compliance with their obligations under Articles 50(2) and (5) of the act and the commitments and measures as specified in Section 1 of the code, signatories will be required to develop, maintain and implement a testing, verification and compliance framework, in line with the state of the art. For example, testing of marking and detection solutions should involve independent experts and/or be

designed in the context of AI regulatory sandboxes under regulatory supervision and should take into account available benchmarks and other measurement and testing methodologies, including those developed or recognised by the AI Office.

To ensure that the marking and detection solutions are future-proof, signatories will be expected to implement "*an adaptive threat modelling approach*". Appropriate training (proportionate to the size and resources of the provider) must be provided to personnel involved in the design and development of AI systems and models and overseeing the compliance, and signatories will cooperate with competent market surveillance authorities to demonstrate compliance with their commitments under the code and provide all relevant information and access to the system.

Section 2: Rules for labelling of deepfakes and AI-generated and manipulated text applicable to deployers of AI systems

Section 2 of the draft code is split into three parts. Part A covers deployers' **general commitments** relating to their obligations under Article 50(4) of the act; Part B describes a specific commitment and measures relating to **deepfakes**; and Part C covers a specific commitment and measures for **AI-generated and manipulated text**.

A: General commitments

- **Commitment 1: Disclosure of origin of AI-generated and manipulated content based on a common taxonomy and an icon** — Code signatories who are deployers of AI systems that generate deepfakes or text publications falling within the scope of Article 50(4) of the act commit to apply consistent disclosure of origin and to use a common taxonomy classifying such content. The taxonomy will distinguish between "fully AI-generated content", i.e., content fully and autonomously generated by the AI system without human authored authentic content (e.g., based solely on prompts), and "AI-assisted content" that involves a mix of human and AI involvement (including relatively minor AI-alterations that change the context of the content, such as noise removal). Signatories will also apply a common icon for deepfakes and AI-generated and manipulated text publications as a method of disclosure. This must be clearly visible at the time of the first exposure and placed in a position appropriate to the content format and dissemination context.
- **Commitment 2: Compliance, training and monitoring** — To effectively fulfil and demonstrate compliance with their AI Act obligations and code commitments and measures, signatories will need to develop, maintain and implement internal compliance and monitoring documentation (proportionate to the size and resources

of the deployer), as well as cooperation mechanisms.

Signatories will facilitate the possibility for third parties and natural persons to flag mislabelled or nonlabelled content. Specifically, signatories commit to cooperate with market surveillance authorities and other third parties who have an interest in determining and/or evaluating whether content has been appropriately labelled (such as media regulators, providers of intermediary services, including Very Large Online Platforms and Very Large Online Search Engines as defined in the Digital Services Act). Signatories will be expected to follow up on reported instances of noncompliance "without undue delay". Appropriate training must be provided to personnel involved in the creation, modification or distribution of content covered by Article 50(4) of the act.

- **Commitment 3: Ensure accessible disclosure for all natural persons** — Signatories commit to ensure icons with associated labels are accessible and conform to applicable accessibility requirements under EU law. To comply with this measure, signatories are encouraged to provide support to implement any available relevant standard, including ETSI EN 301 549 "Accessibility requirements for ICT products and services".

B: Specific commitment and measures relating to deepfakes

- **Commitment 4: Specific measures for deepfake disclosure** — Signatories will set up and implement internal processes to: (i) identify deepfake image, audio, video content; (ii) apply the definition of deepfake in a consistent manner; and (iii) determine whether applicable exceptions apply (e.g., law enforcement use) or if the content relates to artistic, creative, satirical and fictional work.

  Signatories will disclose any deepfake content in a clear and distinguishable manner at the time of the first exposure (if not earlier). How this should be done will be dependent on the type of content. For real-time deepfake video, for example, signatories will display the icon in a nonintrusive way consistently throughout the exposure where feasible. For non-real-time deepfake video, signatories will use the icon that could be combined with a disclaimer at the beginning of the exposure. For deepfake images, the icon should be clearly distinguishable and visible; for audio, an oral disclaimer should be provided at the beginning; and for longer audio formats such as podcasts, at the beginning, at intermediate stages and at the end.

  With regard to deepfake content that forms part of evidently artistic, creative, satirical, fictional or analogous work or programmes, signatories will disclose such deepfakes in an appropriate manner that does not hamper the display or enjoyment of the work, including its normal exploitation and use, while maintaining the utility and quality of the work and appropriate safeguards for the rights and freedoms of third parties. The icon should therefore be placed in a nonintrusive position, again depending on the type of content.

C: Specific commitment and measures for text publications

- **Commitment 5: Specific measures for disclosure of AI-generated or manipulated text** — the draft code specifies measures signatories will commit to implement in order to correctly identify all AI-generated or manipulated text published with the purpose of informing the public on matters of public interest, where no human has reviewed the text publication and no natural or legal person has assumed editorial responsibility (AI-generated and manipulated text publications) and to ensure clear, distinguishable and timely disclosure.

  Signatories will set up and implement internal processes to correctly identify AI-generated or manipulated text publications in a consistent manner and will disclose the AI-generated and manipulated text publications in a fixed, clear and distinguishable manner at the time of the first exposure (if not earlier).

  To rely on the exception relating to human review, editorial control and responsibility, signatories will establish internal procedures and maintain documentation (which are proportionate to the deployer's size) demonstrating that the AI-generated or manipulated text publications have undergone human review or editorial control and that a natural or legal person has editorial responsibility. As an optional step, signatories may record additional information pertaining to the nature of the human review or the type of AI involved.

Comment

The rules covering the transparency of AI-generated content will take effect from 2 August 2026. To meet that deadline, the European Commission intends to have a second draft of the code drawn up by mid-March 2026, with the code expected to be finalised by June. The first draft is therefore very much that, a first draft, and in that regard maps out what the commission refers to as the "high-level" and "key" considerations for providers and deployers of AI systems generating content falling within the scope of Articles 50(2) and (4) of the act.

During that seven-month timeframe, more work will be done to refine the code and to develop other aspects on which specific stakeholder input is sought, in particular technical considerations on feasible approaches to marking AI-generated software code and more novel or challenging kinds of content such as very short texts, as well as audio-only labelling and the evolution of a common icon.

Finally, while the commission refers to the proposed code as a “voluntary tool” to demonstrate compliance, it uses familiar language in describing it as a “*guiding document*”. In-scope content providers and platforms, particularly those operating at scale, are therefore faced with the choice of signing up to a prescriptive code (which is not in itself definitive evidence of compliance) or applying alternative methods and processes. In case of the latter, regulators are likely to judge such alternative measures against the code commitments.

Authors

Emma L. Flett

Partner / London

Max Harris

Partner / London

Prasanth Kapilan

Associate / London

Related Services

Practices

- Technology & IP Transactions

Suggested Reading

- 01 August 2024 Kirkland Alert AI Act Arrives: EU Equips AI With a New Rulebook
- 04 December 2023 Kirkland Alert President Biden Signs Executive Order Introducing New Regulatory Framework for AI-Enabled Technology in Healthcare
- 16 May 2023 Kirkland Alert Washington’s My Health My Data Act

This publication is distributed with the understanding that the author, publisher and distributor of this publication and/or any linked publication are not rendering legal, accounting, or other professional advice or opinions on specific facts or matters and, accordingly, assume no liability whatsoever in connection with its use. Pursuant to applicable rules of professional conduct, portions of this publication may constitute Attorney Advertising.