Second Draft Code of Practice on Transparency of AI-Generated Content

Kalina Bontcheva

Working Group 1 Chair

Dino Pedreschi

Working Group 1 Vice-Chair

Christian Riess

Working Group 1 Vice-Chair

Anja Bechmann

Working Group 2 Chair

Giovanni De Gregorio

Working Group 2 Vice-Chair

Madalina Botan

Working Group 2 Vice-Chair

Table of Contents

Introductory statement by the Chairs and Vice-Chairs

As the Chairs and Vice-Chairs of the two Working Groups, we hereby present the second draft of the Code of Practice on Transparency of AI-generated Content under the AI Act (the "Code") open for stakeholder feedback ***by 30 March 2026 (22:00 CET) through the EUSurvey available to the Code participants and observers****. This second draft of the Code addresses key considerations for providers and deployers of AI systems generating content that falls within the scope of Article 50(2) and (4), identified as part of the work of two Working Groups working in close collaboration:*

- *Working Group 1: Requirements for marking and detection of outputs of generative AI systems (Article 50(2) and (5) AI Act)*
- *Working Group 2: Requirements for disclosure of deep fakes and certain AI-generated text (Article 50(4) and (5) AI Act)*

This second draft incorporates written feedback from hundreds of participants and observers to the Code of Practice process, from industry, academia, civil society and other stakeholders, collected via EUsurvey (open until 23 January 2026) and as part of five meetings and workshops organised in January 2026. This second draft also incorporates contributions from Member States via the AI Board and Members of the European Parliament represented in the IMCO-LIBE Working group monitoring the AI Act implementation. We also received additional feedback from other stakeholders, which we endeavoured to examine within the tight timeline.

The quality of the input received so far has been exceptionally high and we thank stakeholders and Member States for actively and constructively engaging in this process which we trust will lead to a better Code. While consensus might not be possible on all aspects, we have tried to integrate large parts of the relevant feedback and to strike a balance between conflicting views and interests. In carrying out this balancing exercise, we have also tried to stay true to the core objectives of the AI Act transparency obligations and to our task of preparing a Code that facilitates their effective and consistent application.

It is our hope that this second draft of the Code can serve as a sound foundation for the final Code, intended to be a practical, proportionate and useful tool for providers and deployers of generative AI systems of all sizes and sectors to comply with their marking and labelling obligations under the AI Act. Elements that need further development in the current draft, that we aim to address in the final Code, and for which specific stakeholder input is sought, are as follows:

- *For WG1 (Section 1 of the draft Code):*
  - *Technical considerations on the implementation of the revised measures under Commitments 1 and 2.*

    - *Technical considerations on the implementation of Measure 3.4 in Section 1.*
    - *Feedback on the clarity of the terminological definitions in the Glossary of Section 1.*
- *For WG2 (Section 2 of the draft Code):*
    - *Based on input from stakeholders and reviewed research on AI-labelling from academics and industry, WG2 has been working on more general design and placement requirement (Commitment 1, Measure 1.1 and 1.2). Recognizing the fast technological development yet tight release date of the final code, we propose a two-step model for implementing labelling, including in the first step a voluntary, common easy-to-use free label from the EU. We enclose rough first iterations but will further workshop and preferably test icons and additional text explanation (see Measure 1.1) before release together with the final Code.*
    - *In the second step: Stakeholders and research point to promising effects on trustworthiness from an interactive label solution holding a second layer with more details on what has been manipulated. However, state-of-art on such technical solutions will not be available to all deployers. We encourage such solution in Measure 1.1 and we suggest establishing this layer on the common EU icon going forward. For this to happen, we propose the establishment of a taskforce and we are looking forward to the feedback from stakeholders.*

In preparing this second draft, we have been principally guided by the provisions in the AI Act in determining matters that fall within the scope of the Code. Accordingly, unless the context and definitions contained within the Code indicate otherwise, the terms used in the Code should be understood and interpreted as they are in the AI Act. Those elements from the first draft identified as going beyond the scope of the AI Act obligations have been transformed into voluntary measures. Other elements related to the scope of key definitions and exceptions from the obligations have not been addressed in this draft Code since they will instead be covered in Commission guidelines on Article 50 AI Act that are being developed by the Commission in parallel.

Additional time for consultation and deliberation – both externally and internally – will be needed to refine and improve this second draft of the Code. As a group of independent Chairs and Vice-Chairs, we strive to make this process as transparent and accessible to stakeholders as possible, aiming to share our work and our thinking as early as possible while taking sufficient time to coordinate and discuss key questions within Working Groups. We count on your continued engaged collaboration and constructive feedback.

We invite stakeholders to review the document and provide feedback to help shape the third and final version of the Code, which will play a crucial role in facilitating the transparency of AI-generated content in the EU. We welcome written feedback by the

Code of Practice participants and observers ***by 30 March 2026 (22:00 CET) through the EUSurvey available to the Code participants.***

We are very much looking forward to the next stakeholder meetings in March and to the input that we will receive in this second and last iteration before we submit the final Code to the Commission in time to enable effective implementation of the transparency obligations in the AI Act before the rules become applicable on 2 August 2026.

Thank you for your engagement and support!

Kalina Bontcheva
Working Group 1 Chair

Dino Pedreschi
Working Group 1 Co-Chair

Christian Riess
Working Group 1 Vice-Chair

Anja Bechmann
Working Group 2 Chair

Giovanni De Gregorio
Working Group 2 Vice-Chair

Madalina Botan
Working Group 2 Vice-Chair

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)

-

Kalina Bontcheva
Working Group 1 Chair

Dino Pedreschi
Working Group 2 Vice-Chair

Christian Riess
Working Group 1 Vice-Chair

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

Objectives

The overarching objective of this Code of Practice ("Code") is to improve the functioning of the internal market, to promote the uptake of human-centric and trustworthy artificial intelligence ("AI") and to support innovation pursuant to Article 1(1) AI Act, while ensuring a high level of protection of health, safety, and fundamental rights enshrined in the Charter, including democracy, the rule of law, and environmental protection, against harmful effects of AI in the Union.

To achieve this overarching objective, the specific objectives of this Section of the Code are:

a) to serve as a guiding document for demonstrating compliance with the obligations provided for in Article 50(2) and (5) AI Act, while recognising that adherence to the Code does not constitute conclusive evidence of compliance with these obligations;
b) to assist providers of AI systems generating synthetic audio, image, video or text content in complying with their obligations under Article 50(2) and (5) the AI Act and to enable the competent market surveillance authorities to assess such compliance in relation to such providers who choose to rely on the Code to demonstrate compliance with these obligations.

Recitals

Whereas:

a) **Trust in the information ecosystem:** Signatories recognise that AI systems can generate large quantities of synthetic content and that it is becoming increasingly difficult for humans to distinguish AI-generated content from human-authored authentic content, impacting the integrity of and trust in the information ecosystem and raising new risks of misinformation and manipulation at scale, fraud, impersonation and consumer deception. Signatories recognise that transparency is fundamental to fostering trust in and integrity of the ecosystem and to ensuring that AI systems remain reliable and trustworthy.

b) **Multi-layered approach to technical solutions for marking:** Signatories recognise that for generative AI systems, including general-purpose AI systems, no single active marking technique suffices at the time of drafting the Code to meet the four requirements under Article 50(2) AI Act of effectiveness, interoperability, robustness and reliability. This calls for an implementation of a multi-layered approach, widely recognised as the principal way of enabling reasonable compliance with those four requirements in a way that is aligned with emerging state of the art marking approaches, based on practical experience, international scientific reports[1], expert and stakeholder inputs, and emerging standards in the field. Therefore, an appropriate combination of marking techniques is to be applied to meet the four requirements of Article 50(2) AI Act, as far as this is technically feasible given the output modality, and taking into account potential trade-offs in the implementation of the requirements for effectiveness, reliability, robustness and interoperability, as well as the specificities and limitations of various types of content, the costs of implementation, relevant standards, and the evolving technological state of the art. The Code promotes innovation and future-proofness in the fast-evolving technological space and marking techniques by allowing Signatories in the future to rely on alternative techniques, or even a single technique, as long as they can prove compliance with the four requirements in Article 50(2) AI Act and this Section of the Code based on verifiable common benchmarks.

c) **Cooperation along the value chain:** Signatories recognise the need for practical arrangements for making detection mechanisms accessible, as appropriate, facilitating cooperation with other actors along the value chain, and disseminating content or checking its authenticity and provenance to enable the public to effectively identify AI-generated and manipulated content. Signatories who are generative AI model providers play an important role in the value chain and are accordingly encouraged to facilitate compliance of downstream providers of generative AI systems built on those models.

d) **Advancing innovation in marking and detection techniques:** Signatories recognise that determining the most effective technical methods for marking and detection remains an evolving challenge. Signatories recognise that this Section should encourage providers of generative AI systems and underlying models to advance the state of the art in AI marking and detection techniques and related processes and measures. Signatories further recognise that if providers of generative AI systems can demonstrate equal or superior marking and detection techniques in compliance with Article 50(2) AI Act, these techniques should be recognised as advancing the state of the art in AI marking and detection.

e) **Cooperation with other stakeholders:** Signatories recognise that effective, robust, reliable and interoperable technical solutions for marking and detection merit investment of time and resources. They recognise the advantages of collaborative efficiency, e.g. by sharing methods and/or infrastructure and relying on open standards and marking techniques implemented at the model level or provided by other third parties. Signatories further recognise the importance of enabling relevant third parties and users to detect marked content, and of engaging expert or lay representatives of civil society, academia, and other relevant stakeholders in understanding technical solutions. Signatories recognise that such cooperation may involve entering into agreements to share information relevant to technical solutions, while ensuring proportionate protection of sensitive information and compliance with applicable Union law. Signatories further recognise the importance of cooperating with market surveillance authorities and of fostering collaboration between providers of generative AI systems and models, deployers and other users of generative AI systems, online platforms, researchers, civil society and regulatory bodies to address emerging challenges and opportunities in AI content provenance.

f) **Promoting standardisation:** Signatories recognise the need to support and advance open standards and interoperability. They recognise that further efforts will be required for such standards to emerge from international and European standard-setting organisations, considering the implementation challenges and the fast-evolving field. They recognise the importance of a shared infrastructure to distribute costs and set graduated requirements that scale to organisational capacity. In particular, they recognise that content provenance marking standards need to be elaborated further to capture the provenance chain of content authoring, recording creation or modification carried out by an AI system as well as standards on watermarking and other marking methods to ensure interoperability.

g) **Proportionality for Small and medium enterprises ("SMEs") and small mid-cap enterprises ("SMCs").** To account for differences between providers of generative AI systems regarding their size and capacity, simplified ways of compliance for SMEs and SMCs, including startups, should be possible, in a proportionate manner.

Commitments

This Section of the Code applies only to Signatories in so far as they are providers of AI systems generating synthetic audio, image, video or text content ("generative AI systems"), falling within the scope of Article 2 and Article 50(2) AI Act. Without prejudice to the primary responsibility of providers of generative AI systems under Article 50(2) AI Act, providers of generative AI models that are placed on the market independently from AI systems and third party providers of marking solutions may also, on a voluntary basis, adhere to the Code to demonstrate that their marking and detection solutions comply with the requirements of Article 50(2) and (5)AI Act and this Section of the Code.

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

LEGAL TEXT: Article 50(2) and recitals 133 and 135 AI Act

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.

In order to fulfil their obligation under Article 50(2) AI Act to mark in a machine-readable manner the outputs of generative AI systems, including general-purpose AI systems, Signatories commit to implement active marking with regard to the audio, image, video or text content, or any combination thereof, generated or manipulated by the AI system(s) which they place on the market or put into service in the Union.

In order to fulfil this Commitment, Signatories commit to implement the following Measures as applicable to the respective modality and type of content generated or manipulated by their AI system(s).

Measure 1.1: Machine-readable marking techniques

So long as no single marking approach is sufficient, under the state of the art, to comply with the four requirements in Article 50(2) AI Act of effectiveness, interoperability, robustness and reliability, Signatories will implement a multi-layered marking approach to ensure that the outputs of their generative AI systems are marked with at least two layers of machine-readable active marking, as specified in the sub-measures below.

The multi-layered approach specified in this measure does not prevent Signatories from being able to demonstrate compliance in the future with an alternative (possibly single) marking technique, provided that they can prove to the competent market surveillance authorities, based on independently verified benchmarks, that their technique(s) achieve at least the same, if not superior, level of robustness, reliability, effectiveness, and interoperability, as required by Article 50(2) AI Act and Commitment 3 of the Code.

The marking techniques may be implemented at different stages of the value chain (the AI system provider itself or an upstream model provider) and may also be provided by third parties (in particular, providers specialised in transparency marking techniques). Signatories may rely on those third party technical solutions as long as they are provided by organisations that have adhered to this Section of the Code and have demonstrated to the AI Office and the competent market surveillance authorities that those solutions are compliant with the requirements in this Section of the Code. Such reliance is without prejudice to the Signatories' own responsibility under the AI Act and the Code to ensure that the outputs of their generative AI systems are suitably and compliantly marked.

Sub-measure 1.1.1: Digitally signed metadata

If content is generated or exported in a data format that supports adding information as part of the metadata (e.g., an audio, image, video, or document file), Signatories will record and embed through the metadata information whether the content is AI-generated or whether it is AI-manipulated, an interoperable identifier that can be referenced by other layers (e.g., watermark/fingerprint) and information regarding how to access the provider's marking detection tool specified in Measure 2.1.

Signatories will ensure the metadata marking complies with the quality requirements specified in Commitment 3.

All added information will be digitally signed and time-stamped in a secure and tamper-evident manner. Signatories will adopt means for ensuring the secure usage of the signing certificates and the confidentiality of the associated private keys.

In the case of free text and other output types that are not available in a format that hosts metadata, Signatories are encouraged to implement an option that allows the download of a digitally signed manifest containing a certified version of the output generated or manipulated by their AI system to certify the artificially generated or manipulated origin of the text content. Such a provenance certificate will enable deployers and other users to provide third parties with certificates that the content is AI-generated or manipulated, linking it back to the specific generative AI system.

Sub-measure 1.1.2: Imperceptible watermarking techniques interwoven within the content

Signatories will ensure that AI-generated or manipulated content is marked with an imperceptible watermark, with the exception of very short text where applying embedded watermarking techniques fail to meet the reliability requirements specified in Measure 3.2. This watermark will be directly interwoven within the content in a manner that is difficult for it to be separated from the content. Signatories will ensure the watermark complies with the quality requirements specified in Commitment 3.

The watermark will serve as a robust backup of the provenance information provided in the secure metadata under sub-measure 1.1.1, where available.

To protect user privacy, the watermark will be designed so that a fragment of the content suffices to detect the watermark, where feasible and contingent to compliance with the quality criteria in Commitment 3.

Signatories may embed watermarks during AI model training, AI model inference, or within the output of an AI system or its underlying AI model. Signatories who provide generative AI models are encouraged to implement relevant marking techniques at the model level to facilitate compliance of downstream providers of generative AI systems built on those models in a manner that helps them meet the quality requirements specified in Article 50(2) AI Act and in Commitment 3.

Signatories will add an additional publicly readable watermark to audio, image, or video content, indicating the provider for the marking detection tool and an identifier for the metadata. While such a public marking cannot provide security guarantees, it can nevertheless facilitate verification in cases where metadata has been stripped from the content, as further detailed in Measure 3.4.

Signatories will ensure that multimodal output of their AI system is additionally synchronised to the extent technically feasible across the modalities in a manner that the marking is recognisable when only one or a subset of modalities have been altered or exchanged.

Sub-measure 1.1.3: Fingerprinting or logging facilities (optional supplementary measure)

Where appropriate to address deficiencies in the effectiveness, reliability and robustness or other limitations of the marking techniques described in sub-measures 1.1.1 and 1.1.2, and taking into account potential trade-offs related to privacy and security, as well as scalability challenges and costs, Signatories may implement as an optional supplementary measure fingerprinting or logging mechanisms for the AI-generated or manipulated content that allow for checking whether an output has been generated or manipulated by their generative AI system. For example, direct logging might be appropriate for text outputs, whereas perceptual hashing or other fingerprinting approaches may be preferable for audio and visual outputs.

Signatories will ensure the fingerprinting or logging is limited to the output data and is implemented in a secure and privacy-preserving manner in compliance with EU data protection law and respect for media freedom, editorial independence and journalistic source protection obligations. Deployers and other users of the generative AI systems will be given access to the logging policies and granted control over what is logged, how data is stored, and how long it is retained with clear procedures for secure deletion of logs containing output data after its intended purpose has been fulfilled.

This sub-measure shall in no way be interpreted as creating a general commitment to log, monitor, or retain prompts for the AI-generated content or user interactions.

Measure 1.2: Non-removal of machine-readable marking

In addition to the requirement in Article 50(2) AI Act, as detailed in measure 3.3. that requires high robustness of the marking techniques against expected downstream processing and adversarial attacks, Signatories will make best efforts to preserve marks on content generated or manipulated by their AI system by applying the following cumulative measures:

a) Signatories will retain and abstain from altering or removing existing metadata to the extent technically feasible, including where such content is used as input and subsequently transformed by their AI system into a new output, except where content transformation requires updating marks to maintain accurate provenance chain; and

b) Signatories will include in the acceptable use policy, terms and conditions of or the documentation accompanying their generative AI system a prohibition for the intentional removal of or tampering with the marks by deployers or any other third party, unless removal is undertaken for the purpose of benchmarking the security of a marking solution or any content transformations and editorial control are recorded in the provenance chain, where available. For AI systems and models provided under free and open licenses, it is sufficient for Signatories to alert users to this requirement in the documentation accompanying the AI system or the AI model without prejudice to the free and open-source nature of the license.

Measures specified in point (a) above do not imply responsibility of the Signatory for third-party markings. Enabling detection of those marks remains the responsibility of the original provider of the generative AI system.

Signatories who operate an online platform or search engine or who otherwise disseminate content to the public are encouraged to ensure that the platform or the search engine preserves metadata and other marks for AI-generated or manipulated content.

Measure 1.3: Transparency of the provenance chain (optional)

Signatories are encouraged to apply provenance standards providing further information about the provenance chain of AI-generated or manipulated content across workflows where technically feasible for the specific modality.

In addition to the information recorded in the metadata pursuant to measure 1.1.1, Signatories may add or record other relevant content provenance information within their AI systems in a way that distinguishes the additional operation(s) performed by their AI system from previous operations, by leveraging metadata to record and verify the provenance chain where technically feasible.

The other provenance information that Signatories are encouraged to record includes the AI system and underlying model identifier, version number, company name of the AI provider, and a timestamp indicating when the content was generated or manipulated. For AI-manipulated content, it is recommended that the metadata contains information about the type of the operation performed by the AI system to modify the content (e.g., object removal). Multiple discrete processing steps carried out by the AI system are recommended to be encoded into a single marker to constrain complexity and to reduce burden.

Where a human carries out an operation in AI-human workflows, the only information that is recommended to be recorded is that a human carried out a given operation (e.g. editing). In such cases, the human author may also encode on a voluntary basis other descriptive information, such as the organisation name and copyright, as applicable.

Measure 1.4: Optional functionality for perceptible markings (for deep fakes and AI-generated and manipulated published text)

In order to facilitate compliance of deployers of generative AI systems with their obligation to disclose deep fakes and certain AI-generated and manipulated published text pursuant to Article 50(4) AI Act, Signatories who are providers of generative AI systems that are capable of generating or manipulating such content are encouraged to provide an optional functionality in their system's interface and implement an integrated option that allows deployers and other users to directly – upon generation of the output – apply at their own discretion a perceptible machine-readable mark or label.

Signatories are encouraged to implement such a functionality for perceptible marks and/or labels in consistency with the Commitments and Measures in Section 2 of the Code. It is also recommended that Signatories follow harmonised UX standards, to the extent possible, in an interoperable manner with existing standardised content management systems and workflows of media publishers and online platforms.

Signatories are also encouraged to implement other supporting measures for displaying labels and provenance metadata that enable deployers and providers of online platforms and websites to implement display practices and policies that are appropriate for their use cases.

This measure is without prejudice to the responsibility of deployers who remain responsible for the disclosure of deep fakes and AI-generated or manipulated published text in a clear and distinguishable manner in accordance with Article 50(4) and (5) AI Act.

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

LEGAL TEXT**:** Article 50(2) and 50(5) and recitals 133 and 135 AI Act

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* […] *detectable as artificially generated or manipulated.*

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.*

In order to fulfil their obligations under Article 50(2) and (5) AI Act to ensure that the outputs of their AI system(s) are detectable as AI-generated or manipulated, Signatories commit to implement the following measures to enable the detection of audio, image, video or text content, or a combination thereof, as generated or manipulated by their AI system and to ensure this information is provided to natural persons concerned in a clear, distinguishable and accessible manner through tools or APIs, and ideally also as forensic detectors.

Measure 2.1: Detection mechanisms for active marking made available to deployers, end-users and other third parties

Signatories will ensure that an interface is made available free of charge to enable deployers, users of the generative AI system, end-users exposed to the content, and other legitimate parties (such as competent authorities, independent researchers, civil society and media organisations) to verify whether content has been generated or manipulated by their AI system based on the active marking techniques specified in Commitment 1. This interface may be an API or a user interface. Alternatively, Signatories will ensure that there is a publicly available detection tool for the content generated or manipulated by their AI systems**.** For privacy and security reasons, Signatories will ensure that the publicly available detection tools or the detection mechanism and the interface is implementable locally or hosted within the European Union and compliant with EU data protection law.

In order to meet the requirements of this measure, a Signatory who has adopted a marking standard or a third party marking technique can rely on the publicly available detection tool(s) provided for the same standard or the third party marking technique in so far as the detection tool(s) comply with Article 50(2) and (5) AI Act and this Section of the Code. Signatories who are providers of generative AI models and who have implemented marking techniques at the level of the model are encouraged to enable downstream providers who build generative AI systems on those models to rely on the model detection mechanism for the content generated or manipulated by their AI system and underlying model in a manner that helps them comply with Article 50(2) and (5) AI Act and this Section of the Code.

Signatories will ensure that the detection mechanism(s) comply with the quality criteria in Commitment 3. Signatories will ensure that the detection mechanism(s) are interoperable, as required in Measure 3.4. Provenance information that is in the metadata from AI system providers other than the Signatory will also be disclosed in the detection interface to the extent that this information can be recovered from the provenance chain. In case a publicly available detector is released, Signatories will ensure that full support and cooperation is provided to the Commission and the market surveillance authorities for the deployment, use and ulterior customization and development of the mechanism.

Signatories are also encouraged to collaborate with the Commission and other relevant actors to make their detection mechanism(s) directly available in distribution and communication platforms and to maintain the mechanism(s) throughout the AI system's lifecycle.

If the detection mechanism requires uploading content, Signatories will ensure that the content is handled as potentially privacy sensitive. In particular, Signatories will not retain a verbatim copy of the content or personally identifiable information about the person requesting verification. This does not exclude privacy-preserving fingerprinting or logging of the content subject to verification pursuant to Measure 1.1.3.

Signatories will ensure that the detection and verification results are provided in accordance with Measure 3.4 below and can be downloadable upon request in a digitally signed document for later reference.

Where a Signatory has developed its own detection mechanism and goes out of business or stops providing that mechanism for other reasons, it will make that mechanism available to the competent market surveillance authorities to ensure legacy content generated or manipulated by their AI system or model remains detectable.

Measure 2.2: Forensic detection mechanisms (optional supplementary measure)

Signatories are encouraged to collaborate with the Commission and competent market surveillance authorities and, as appropriate, research organisations and other relevant stakeholders, to support the development of forensic detector(s) capable of detecting the outputs of generative AI models available on the Union market, including when integrated into systems.

To complement the marking techniques specified in Commitment 1 and contingent upon their capacities, Signatories may implement, as an additional optional measure, forensic detection mechanisms to detect content generated or manipulated by their AI system or underlying model for which marking has been stripped or for which active marking for the content was not technically feasible. Signatories will ensure that the forensic detection mechanism complies with the quality criteria in Commitment 3.

Measure 2.3: Clear and accessible disclosure of verification and detection results

In order to fulfil their obligation under Article 50(5) AI Act, Signatories will ensure that the detection and verification results are presented in a way that is clear and thus easily comprehensible to laypersons who want to verify the origin of the content.

Signatories will ensure the detection and verification results provide information whether they are based on a watermark, metadata or forensic detection or other techniques, to the extent technically feasible.

That information will include the provider of the mark and any additional information that can be detected from the markings (for example, whether the content was AI-generated or AI-manipulated). In addition, Signatories will, to the extent technically feasible, provide an indication of the confidence in the correctness of the detection results.

Signatories will ensure that the results of the detection mechanisms and, where applicable, user interfaces, are accessible to persons with disabilities, in compliance with applicable accessibility requirements under Union law, and in particular with the European Accessibility Act (EAA)1 and the Web Accessibility Directive2. Signatories are encouraged to implement any available relevant standard, including but not limited to the harmonised standard ETSI EN 301 549 “Accessibility requirements for ICT products and services”, and WCAG 2.1 Level AA “Web Content Accessibility Guidelines”.

Measure 2.4: Support literacy on AI marking technologies and verification

Signatories are encouraged to ensure that layperson-oriented documentation and other relevant information (excluding trade secrets) is provided to deployers and other users to support them in making informed decisions on what marking and detection mechanism(s) they may use, including helping them to understand how to access and apply detection mechanisms and to interpret the provenance data and the detection results.

In addition to deployer-focused materials, Signatories are also encouraged to ensure that end-user literacy resources are provided, as appropriate, and calibrated to end-user needs where the AI systems serve populations with lower AI literacy or in sensitive contexts (e.g. educational contexts, youth or elderly users).

These materials may either be developed by the Signatories themselves or created jointly through efforts involving other providers or by organisations or initiatives they belong to. This measure should be implemented in a proportionate manner, taking into account the level of awareness of the deployers and other users of the generative AI system and end-users of the content, the size and resources of the provider, in particular with regard to SMEs and SMCs.

Signatories are encouraged to collaborate with academia, civil society, media and other relevant organisations to promote literacy and awareness regarding AI content provenance and verification, and to support EU-level initiatives to foster consistent understanding of provenance and detection across Member States.

Commitment 3: Measures to meet the Requirements for Marking and Detection Techniques

LEGAL TEXT: Article 50(2) and recital 133 AI Act

2. […] *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.*

In order to fulfil their obligation under Article 50(2) AI Act to ensure the employed technical solutions for marking and the detection of AI-generated or manipulated content are effective, interoperable, robust and reliable, as far as this is technically feasible and taking into account the specificities and limitations of various types of content, the costs of implementation and the generally acknowledged state of the art, Signatories commit to ensure compliance with these requirements in a balanced manner and ensuring that all requirements are met, as outlined in the following measures.

Signatories commit to strive to achieve the highest possible effectiveness, interoperability, robustness and reliability of the marking and detection solutions as described in Measures 3.1. 2, 3.3 and 3.4 below. These requirements shall be achieved to the extent technically feasible, taking into account specificities and limitations of various types of content, aligned with the state of the art, and moderated by hard operational constraints. Operational constraints may impose limitations on computational time and resources, costs of implementation and on scalability to very large or very small content. For example, the watermark in a live video stream must be inserted in real-time, and hence the reliability and robustness requirements can only meet expectations for the state of the art in real-time watermarking. However, operational constraints cannot be used to circumvent compliance with the marking and detection measures in Commitments 1 and 2.

Signatories commit to implement the measures and meet the requirements as specified below prior to placing their generative AI system on the market or putting it into service, and throughout its lifecycle, striving to ensure alignment with the evolving state of the art.

Signatories commit to assess and demonstrate compliance with the measures and requirements specified in this Commitment in accordance with the verification, testing and compliance processes as specified in Commitment 4.

Measure 3.1: Effectiveness

Signatories will implement technical marking and detection solutions which, in conjunction, are fit-for-purpose and capable of effectively enabling natural persons to distinguish between artificially generated or manipulated content and human-authored content, and which contribute to the trust and integrity of the information ecosystem.

To be effective, those solutions should ensure that natural persons can access available disclosures and detection mechanism(s), understand the meaning of detection results, and use those results to make informed decisions, thereby reducing the risk of deception and manipulation and supporting trust and the integrity of the information ecosystem.

Signatories will also demonstrate that their marking and detection solutions contribute to improved trust and integrity of the information ecosystem by evaluating how freely and readily accessible the detection mechanisms are for competent authorities and key organisations working towards checking the authenticity of content and ensuring the integrity of the information ecosystem (e.g., media, fact-checkers, trusted flaggers, independent researchers, NGOs, and other relevant stakeholders).

Measure 3.2: Reliability

Signatories will implement marking and detection solutions that achieve a high level of reliability in different expected contexts and across use cases, to the extent technically feasible and aligned with the state of the art.

The reliability of each marking, and corresponding detection, solution applied indicates how well it distinguishes AI-generated or manipulated content from human-written content. Reliability should be considered to have two components: (i) how accurate the detection of the marking is under controlled conditions, and (ii) how accuracy of the marking and detection solutions vary with respect to the length, entropy and semantics of the content.

Signatories will measure the accuracy of the detection of AI-generated or manipulated content by using relevant state-of-the-art metrics, such as false positive rate and false negative rates of the detection and bit error rates in decoded marker information (if applicable). Low false-positive and false-negative rates will be demonstrated on samples of AI-generated and unmarked human-authored content unseen during the training and development of the AI system and its underlying model.

Signatories will ensure these measurements are performed as a function of the length and entropy of the content generated by the AI system. These measures will be performed in accordance with the type and modality of content the system is intended to generate. For general-purpose AI systems, Signatories will also ensure these measurements are performed on content with diverse semantics, i.e. across varying application contexts and use cases, in particular to demonstrate that the implemented marking and detection solutions generalise across diverse AI-generated or manipulated content per modality (e.g. images such as people, landscapes, or food).

Measure 3.3: Robustness

Signatories will implement marking and detection solutions that achieve a high level of robustness to common alterations and to adversarial attacks, to the extent technically feasible and in a manner that is aligned with the state of the art.

First, Signatories will ensure their marking and detection techniques are robust to typical processing operations. Such typical processing operations are for example mirroring, cropping, compression, screen capturing, paraphrasing, character deletions, changes in image or video resolution, pitch shifting, time stretching, desynchronization, noise removal, voice enhancement or change of format. Signatories will use the same performance metrics as for reliability in Measure 3.2 to assess the robustness of the outcome of such typical processing operations.

Second, Signatories will ensure that adversarial robustness of their marking and detection solutions is assessed in terms of resilience to adversarial attacks such as copying, removal, regeneration, modification, and amortisation attacks on the markings. The robustness requirement to adversarial attacks does not apply to the second public watermark in Measure 1.1.2.

Signatories will measure adversarial robustness of their marking and detection solution by the ability to verify marking integrity, i.e. whether a mark has been tampered with or modified to misrepresent the content's origin. The considered attacks shall be chosen as plausible real-world threats based on the type of content and the type of mark.

Signatories will use attack success rate as a performance metric for adversarial robustness in combination with the performance metrics for reliability in Measure 3.2 to verify that the performance in absence of an attack is not excessively degraded.

Signatories will apply standard cybersecurity practices such as rate limits to prevent and counteract malicious use and attacks against the marking and detection mechanisms. Signatories are encouraged to frequently update their threat assessment to keep up to date with changes in the threat landscape.

Measure 3.4: Interoperability

Signatories will implement technical solutions for the marking and detection of AI-generated or manipulated content that work, to the extent technically feasible, across distribution channels and technological environments, regardless of the application domain, context or content modality formats.

The aim is to ensure full interoperability of the marking and detection solutions of different providers of AI systems with common marking and detection standards. As the state of the art matures and as uniform international and European standard(s) for the marking and detection of AI-generated and manipulated content emerge, interoperability may be implemented in a staged approach, aligned with the state of the art and available standards for different marking and detection techniques. At this stage, with regard to metadata, Signatories will implement or adopt a secure, tamper-evident, open verification standard for signed provenance manifests and digital signature with an interoperable identifier for the other marking layer(s).

Furthermore, Signatories will encode in the second, imperceptible public mark from Measure 1.1.2 information on how to access the corresponding verifier (e.g., in the form of a watermark vendor ID).

Signatories will cooperate with the Commission to create a shared repository of the i) public watermarks for their marking solutions, ii) addresses of the metadata repository including information to verify and interpret the cryptographically secured metadata and iii) the addresses for the detector of the other secure watermark. In case that Signatories implement asymmetric watermarking techniques compliant with the other measures in this Commitment, then they will share the information necessary for detection. . Signatories will further ensure the detection methods they employ are made available in accordance with a standardised open detection protocol that may be developed to detect outputs of generative AI systems within the scope of Article 50(2) AI Act.

Furthermore, Signatories will cooperate in the development of an interoperable provider-agnostic detection interface that can be executed locally on a computer and that can provide a common entry point to all detection mechanisms employed by providers of generative AI systems. Signatories will cooperate with the Commission to ensure the shared provider-agnostic detection interface is connected and provides access to their detection mechanisms including for watermarks that cannot be locally detected.

The feasibility and the potential implementation of such an EU-wide shared provider agnostic detection interface is to be further assessed by the Commission in consultation with the AI Board. Should such an EU-wide detection interface be established, it will incorporate appropriate safeguards, including for privacy and confidentiality protections for proprietary algorithms, security controls preventing reverse engineering, and clear governance structures defining access rights and liability allocation.

Signatories are encouraged to join and/or support international and European standardisation organisations or fora and consortia initiatives focused on the development of content marking and detection standards that operationalise the measures envisaged in this Section of the Code, in particular content provenance standards, as well as watermarking standards allowing for controlled renewal, revocation, or replacement, without degrading any underlying intrinsic provenance signals.

Signatories, including SMEs and SMCs, are encouraged to make use of relevant content marking standards that emerge from international and European standardisation organisations and widely adopted open technical standards that are in compliance with the Code and Article 50(2) and (5) AI Act to promote interoperability and broad adoption and to minimise costs of compliance.

Measure 3.5: Advancing the state of the art in marking and detection

Contingent upon their capacity and resources, Signatories are encouraged to invest in scientific research and development and collaborate with competent authorities, researchers, civil society organisations and other relevant stakeholders to advance the state of the art in marking and detection mechanisms for AI-generated and manipulated content. Signatories are encouraged to participate in a taskforce facilitated by the AI Office to hold regular dialogue meetings with these stakeholders on taking stock and advancing marking and detection techniques for the transparency of AI-generated and manipulated content.

Furthermore, Signatories are encouraged to cooperate on the development of effective, interoperable, robust and reliable and of watermark schemes that enable controlled renewal, revocation, or replacement without degrading the quality of the original output, the development of future forensic models and fingerprinting techniques, as well as the development of shared benchmarks and red-teaming exercises to accelerate progress, while maintaining security.

Commitment 4: Testing, verification and compliance

LEGAL TEXT: Article 50(2) and 50(5) and recital 133 AI Act

In order to effectively fulfil and demonstrate compliance with their obligations under Article 50(2) and (5) AI Act, as well as with the Commitments and Measures specified in this Section of the Code, Signatories commit to set up, keep up to date and implement testing, verification and compliance processes, as specified in the following measures.

Measure 4.1: Compliance framework

Signatories will draw up, implement, and update, in line with the state of the art, a compliance framework that outlines the marking and detection processes and the measures that the Signatories implement to ensure compliance with Article 50(2) and (5) AI Act and the Commitments and Measures in this Section.

The framework will contain a high-level description of implemented and planned processes and of measures to adhere to this Section of the Code and to maintain and keep up to date relevant documentation to be shared with competent market surveillance authorities upon request. This measure should be implemented in a proportionate manner, taking into account the size and resources of the provider, in particular with regard to SMEs and SMCs. Signatories can demonstrate compliance through existing processes and compliance frameworks to the extent that they fulfil the measures in this Section of the Code.

Where Signatories rely on marking and detection solutions provided by third parties or implemented at the level of the generative AI model, Signatories will employ solutions for which those parties adhere to the Code and have demonstrated compliance with this Section of the Code and Article 50(2) and (5) AI Act. This possibility of reliance is without prejudice to the ultimate responsibility of the Signatory as a provider of the generative AI system to ensure compliance with Article 50(2) and (5) AI Act.

Measure 4.2: Testing, verification and monitoring

Prior to the placement on the market and regularly thereafter, Signatories will test the marking and detection solutions for their compliance with the requirements and the measures specified in this Section of the Code in real-world conditions. Signatories who are downstream providers of generative AI systems may rely on results of testing performed by an upstream model or a third party provider of marking and detection techniques, as long as they comply with the requirements specified in this measure.

In the context of testing and evaluation, Signatories will take into account available state-of-the-art benchmarks and other measurement and testing methodologies, including benchmarks and frameworks developed or recognised by the AI Office in collaboration with the AI Board. Such benchmarks should be updated in accordance with the state of the art and reflect realistic transformations and adversarial scenarios. Signatories may involve independent experts in the testing or conduct such testing and evaluation under regulatory supervision in the context of AI regulatory sandboxes as provided for in Article 57 AI Act.

To ensure that marking and detection solutions are future-proof, Signatories will implement an adaptive threat modelling approach, moving beyond generic robustness benchmarks by defining realistic and use-case specific threat scenarios (e.g., recompression, transcoding, speech-to-speech revoicing) to support the development of adaptive defence mechanisms. They will also track real-world degradations, periodically re-evaluate detection thresholds and update detection mechanism(s) to keep false positive rates low, while preserving detectability.

Signatories will implement and document appropriate follow-up corrective actions on compliance shortcomings reported by deployers, independent researchers, civil society and other third parties and observed or reported adversarial attacks.

Measure 4.3: Training

Signatories will make proportionate efforts to provide appropriate training to personnel with roles relevant to ensuring compliance with Article 50(2) and (5) AI Act who are involved in the design and development of AI systems and models and who are responsible for ensuring that the measures specified in this Section of the Code are effectively implemented. This measure should be implemented in a proportionate manner, taking into account the size and resources of the provider, in particular with regard to SMEs and SMCs.

Measure 4.4: Cooperation with market surveillance authorities

Signatories will cooperate with competent market surveillance authorities under the AI Act to demonstrate compliance with Article 50(2) and (5) AI Act and their commitments under this Section of the Code and, at their reasoned request, provide all relevant information and access to the system. Access will be provided via a secure channel and technical interfaces for cooperation, including access to legal-grade detectors and certified provenance data. Article 78 AI Act applies to information obtained in the course of market surveillance activities, ensuring trade secrets and confidential information are preserved in accordance with the AI Act and applicable Union law.

Signatories are encouraged to cooperate with the AI Office and the AI Board in providing information about technologies that could be used to provide such transparency across the value chain and that have been assessed to be compliant with the Code, for instance via a repository of available recognised standards and technologies.

Glossary

Wherever this Section refers to a term defined in Article 3 AI Act, the AI Act definition applies. The following terms with their stated meanings are used in this Section of the Code. Unless otherwise stated, all grammatical variations of the terms defined in this Glossary shall be deemed to be covered by the relevant definition.

Term	Definition
Active marking	Addition or embedding of a marking to AI-generated or manipulated content such as a watermark or attached information such as a secure metadata entry. The purpose of this addition is to facilitate detection of this marking and provenance attribution of the AI-generated or manipulated content.
Detection mechanism for active marking	Detection of markings such as watermarks or verification of secure metadata markers that have been purposefully added by a provider of an AI system or a third party (e.g. model provider).
Adaptive threat modelling approach	A defensive measure in cybersecurity to continuously monitor and, if necessary, to adapt the security of a system.
Amortization attacks	A method in which an attacker performs one difficult or time-consuming task upfront and then re-uses that work to make many follow-up attacks much cheaper and faster.
API	Stands for Application Programming Interface, a machine-usable interface to an AI system or another software service from an AI system provider.
Digital signature	A cryptographic signature that enables secure verification of authenticity of the provider and integrity of the signed content
Fine-tuning	The process of further training an already pre-trained, general-purpose AI model on a specialized dataset to optimize its performance for specific tasks, domains, or styles.
Fingerprinting	Detection technique for image, video, audio, or text, based on either hashing or logging.
Forensic detection	Detection of AI-generated or manipulated content which does not depend on the presence of active AI marking. For example, a forensic method may attribute an image to an AI image generator using a signal characteristic in the image data or a machine learning model trained to distinguish AI-generated images from human generated content.
Hashing / Perceptual Hashing	Reduction of audio or visual content to a short identifier for indexing. Can be used for a fast lookup for known content, i.e., a repository of hashes can be queried to find out whether content is known to have been AI-generated or manipulated.
Logging	Verbatim recording and indexing of content (usually text). Can be used for a fast lookup of known content, i.e., a repository of logged entries

	can be queried to find out whether content is known to have been AI-generated or manipulated.
Manifest	A digital document in a format that permits the attachment of metadata.
Provenance Information	A digital record for a piece of content generated or manipulated by an AI system that shows its origin, how and when the content was generated or manipulated and processing applied to the content.
Shared verifier	A detector or verifier for markings originating from multiple providers of AI systems or models that generate or manipulate content.
Synchronization of markings	Cross-referencing between markings in multimodal content. For example, a document consisting of a text and an image may contain a marking in the text that refers to the image, and a marking in an image that refers to the text, such that one cannot replace only the text or only the image without this affecting the integrity of the markings.
User	Either a deployer within the meaning of Article 3 (4) the AI Act or another person that is using the generative AI system of a provider.
UX	"User Experience", i.e., the way how a user interacts with and perceives user-sided aspects of a software product
End-user	A natural person exposed to the content generated by the AI system.
Watermark	A marker directly connected and interwoven within the content, typically through an imperceptible modification of the content, such that it is difficult to remove without affecting the fidelity of the content.

Section 2:

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

Anja Bechmann
Working Group 2 Chair

Giovanni De Gregorio
Working Group 2 Vice-Chair

Madalina Botan
Working Group 2 Vice-Chair

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

Objectives

The overarching objective of this Code of Practice ("Code") is to improve the functioning of the internal market, to promote the uptake of human-centric and trustworthy artificial intelligence ("AI"), while ensuring a high level of protection of health, safety, and fundamental rights enshrined in the Charter, including democracy, the rule of law, and environmental protection, against harmful effects of AI in the Union, and to support innovation pursuant to Article 1(1) AI Act.

The objectives of this Section of the Code are:

a) to serve as a guiding document for demonstrating compliance with the obligations of deployers of generative AI systems provided for in Article 50(4) and (5) AI Act, while recognising that adherence to the Code does not constitute conclusive evidence of compliance with these obligations under the AI Act;

b) to ensure that deployers of AI systems that generate or manipulate image, audio or video content constituting a deep fake or text intended to inform the public on matters of public interest comply with their obligations under Article 50(4) and (5) AI Act, and to enable the competent market surveillance authorities to assess compliance of deployers who choose to rely on the Code to demonstrate compliance with those obligations under the AI Act.

Recitals

Whereas:

a) **Detection and disclosure:** Signatories acknowledge that technological advances in generative AI systems can enhance the realism and persuasiveness of AI-generated or manipulated content, increasing the importance of transparency mechanisms to safeguard public trust and democratic discourse. AI systems capable of generating or manipulating image, audio or video content that appreciably resembles existing persons, objects, places, entities or events may produce content which falsely appears authentic or truthful, raising specific risks for individuals, the integrity of the information ecosystem and democracy. Moreover, AI systems capable of generating or manipulating text that is published with the purpose of informing the public on matters of public interest should also be disclosed to natural persons. Clear and distinguishable disclosure of the artificial origin or manipulation of such content is a necessary safeguard to mitigate the risk of deception and reputational harm and to uphold trust as a public interest.

b) **Clear and user-friendly labelling:** Signatories acknowledge that as deployers of AI systems generating or manipulating deep fakes and AI generated or manipulated text falling within the scope of Article 50(4) AI Act, they are responsible for labelling the output accordingly and for disclosing its artificial origin or manipulation in a manner that is appropriate to the type of modality and context of dissemination. These responsibilities are complementary to the technical solutions implemented by providers under Article 50 (2) AI Act, contributing to increased transparency and trust along the AI value chain. Transparency measures should be user-friendly across the Union to strengthen the ability of the public to distinguish AI-generated or manipulated content and to support the resilience of the information ecosystem.

c) **Artistic creation:** Signatories emphasise that, where the AI-generated or manipulated deep fake content forms part of an evidently artistic, creative, satirical, fictional or analogous work, transparency requirements apply in a proportionate manner. The disclosure of the existence of such AI-generated or manipulated deep fake content should therefore be implemented in a way that does not hamper the display, enjoyment, normal exploitation or creative quality of the work, while preserving appropriate safeguards for the rights and freedoms of third parties as enshrined in the Charter.

d) **Accessibility:** Signatories emphasise the relevance of ensuring accessible disclosure to end-users exposed to the content, particularly in relation to different needs and vulnerabilities. Icons, labels and disclaimers should be designed in a way that ensures they are easily perceivable and understandable by persons with disabilities. This includes, for instance, providing alternative text for screen readers, audio disclosures for visually impaired users, sign language or captioned disclosures for hearing-impaired users, and ensuring sufficient colour contrast and readability.

e) **AI literacy:** Signatories recognise that clear disclosure of AI-generated or manipulated deep fake content and AI generated text publications of public interest within the scope of Article 50(4) AI Act is essential for individual awareness and for supporting AI literacy. Public awareness and clear labelling of such AI-generated or manipulated content can further strengthen individuals' ability to distinguish synthetic content, thereby enhancing the practical impact of the transparency measures set out by this Code.

f) **Additional safeguards under other Union and national law:** Signatories acknowledge that transparency obligations apply alongside, and do not replace, other legal responsibilities that may apply to the creation, distribution or use of AI-generated or manipulated content under applicable Union legislation on data protection, consumer protection, digital services (Digital Services Act[1]), intellectual property, media law

[1] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance), PE/30/2022/REV/1, *OJ L 277, 27.10.2022, pp. 1–102.*

(Audiovisual Media Services Directive[2] and European Media Freedom Act[3]), political advertising, criminal law and other relevant regulatory frameworks.

Commitments

This Section of the Code applies only to Signatories in so far as they are deployers of AI systems that generate or manipulate deep fakes or published text with the purpose of informing the public on matters of public interest falling within the scope of Article 50(4) AI Act. Each reference below to "content" implies content that qualifies as a deep fake under Article 3(60) AI Act (referred to as 'deepfake') or, respectively, text published with the purpose of informing the public on matters of public interest, without human review or editorial control and where no natural or legal person holds editorial responsibility for the publication of the content (referred to as 'published text' or 'text published on matters of public interest').

Commitment 1: Disclosure of AI-Generated and Manipulated Deep Fakes and Published Text

LEGAL TEXT: Article 50(4) and 50(5) and recitals 133 and 135 AI Act

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.

In order to fulfil their obligations under Article 50(4) and (5) AI Act, Signatories commit to ensure consistent disclosure of the artificial origin of AI-generated or manipulated deep fakes or

[2] Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) in view of changing market realities, PE/33/2018/REV/1, OJ L 303, 28.11.2018, pp. 69–92.

[3] Regulation (EU) 2024/1083 of the European Parliament and of the Council of 11 April 2024 establishing a common framework for media services in the internal market and amending Directive 2010/13/EU (European Media Freedom Act) (Text with EEA relevance), PE/4/2024/REV/1, *OJ L, 2024/1083, 17.4.2024.*

published text on matters of public interest by using the uniform EU icon (once available) or choosing an alternative icon or labelling solution that follows the design and placement requirements specified in the following measures.

The disclosure and labelling process may be integrated into existing disclosure and labelling practices of the Signatory to the extent that compliance with this Section of the Code and Article 50(4) and (5) AI Act is ensured.

Signatories recognise that the use of labelling does not exempt them from other Union and Member States' laws, such as those related to the protection of third parties' rights and the fundamental freedoms, including applicable legal requirements with regard to obtaining consent of the depicted person or rightsholders.

Measure 1.1 Design requirements for icons, labels or disclaimers

For content to which a visual icon or label can be applied, Signatories will implement the following design requirements:

The icon or label will hold as the main visual element the capitalized acronym "AI" in the English language (e.g., an AI icon), possibly supplemented, where appropriate, with a short text label regarding the type of involvement of AI (e.g. "Generated with AI", "Made by AI" or "Manipulated with AI"). Where technically feasible and available this can further be elaborated in a second layer detailing (e.g., what has been modified).

- The letters in both the acronym and the text will have the same vertical dimension;
- The icon or label can appear in different styles (e.g., colour and typography), as long as it remains clear, accessible, and distinguishable, i.e. readable and recognisable to all, including vulnerable categories of end-users that may be exposed to the content, such as children, elderly people, and persons with disabilities;
- The icon or label may appear in a different size depending on the context, while ensuring that it is clear and distinguishable. If resized, the proportions of the letters must be preserved;
- The contrast ratio will be maintained of at least 4:5:1 against the background.

For audio-only content, Signatories will include, within the audio content itself, a short audible disclaimer, sound or signal in plain and simple natural language, in the same language as the content (where applicable), disclosing the artificial origin of the audio. Where appropriate, information regarding the type of involvement of the AI system may be provided (e.g. "Generated with AI", "Made by AI" or "Manipulated with AI").

When applying the design requirements of this measure, Signatories will consider the potentially diverse composition of the audience exposed to the content (including diverging levels of AI and digital literacy, language proficiency or general knowledge) and the potential sensitive nature of the context in which the content is used (e.g. in the financial, medical, education or other high-risk sectors).

Signatories will ensure accessible disclosure in different modalities in accordance with applicable Union law, including but not limited to:

- audio descriptions or alternative cues for visual disclosure elements;
- tactile cues for audio-only content, when the device used allows for such cues (e.g. a vibration alert before play), taking into account the needs of end-users with hearing impediments;

- high contrast icons or labels and screen-reader compatibility, including for end-users with colour vision deficiencies;
- detectability of the icon or the label by assistive technologies.

Signatories are encouraged to implement any available relevant accessibility standard or guideline, including but not limited to the harmonised standard ETSI EN 301 549 "Accessibility requirements for ICT products and services" or the W3C Web Content Accessibility Guidelines 2.1, to the extent the Signatories' services or products fall within the scope of such standards or guidelines.

The appendix provides sample icons and labels to be further discussed with the Code of Practice participants in order to support the development of a uniform EU icon that may be used by deployers in the execution of this measure. Once finalised, the EU icon will be made available under the Europe Union Public Licence and will be made available for download on an EU website.

Measure 1.2 Placement requirements for icons, labels or disclaimers

To meet the legal requirements of labelling in a clear and distinguishable manner at the latest at the time of first exposure under Article 50(5) AI Act, Signatories will display the icon, label or disclaimer in an appropriate and perceivable position, in accordance with the content format and dissemination context, taking into account the following overarching principles applicable to all content modalities:

- The icon, label or disclaimer should be affixed on or directly embedded into the AI-generated or manipulated content.
- Where technically feasible, signatories will ensure that the icon, label or disclaimer always travels with the content to which it was applied. Deployers will collaborate on a best effort basis with actors whose services or products are used to further distribute or disseminate the content (e.g. publishers, online platforms or retail) to preserve applied icons, label or disclaimers consistently (including those applied in accordance with Commitment 3).
- Icons, labels or disclaimers should be clearly perceivable at the latest at the time of first exposure of a natural person to the content.
- For deep fake content that is part of artistic, creative, satirical, fictional or analogous works, the requirements detailed in Commitment 3 apply.

Further specifications per type of modality are provided below.

Real-Time Video (multiple modalities)

For real-time deep fake video (including live television broadcasts or livestreaming), Signatories will display an icon or an alternative label consistently throughout the exposure where feasible. In the case where disclosure is done through audio, it should be presented simultaneously with the icon or label.

Alternatively, Signatories can use a visual or audio disclaimer at the latest at the beginning and at regular intervals during the exposure that discloses that the content includes deep fakes. Such a disclaimer should be displayed or broadcasted for an appropriate duration to ensure perceivability.

Non-Real-Time Video (multiple modalities)

For non-real-time deep fake video, Signatories will disclose that the video contains deep fakes with an icon or label. The Signatories may choose among the following disclosure options, individually or combined, as appropriate to the context:

- Long videos: icon or label at the latest at the beginning and repeated at regular intervals (e.g., when the specific deep fake content is shown and/or after commercial breaks).
- Short videos: icon or label consistently throughout the exposure from the beginning of the exposure. If the content of the video is entirely AI-generated and manipulated, this must be indicated throughout. The icon or label should clearly stand out and not be hidden (e.g. it should stand out from the background of the video and not be too close to other overlaying text and icons).
- In both cases: where disclosure is done through audio, it should be presented simultaneously with the visual icon or label.
- A disclaimer in the credits at the end of the video can be inserted. This measure always needs to be accompanied by one or more of the previous options.

Image (single modality)

Recognizing the cross-platform and cross-media transferability of deep fake images, Signatories will place an icon or label consistently at the latest at the first exposure and at any subsequent exposure to the image. The icon or label should be clearly distinguishable, particularly from the image itself, prominently visible, and not obscured or hidden (e.g. embedded within image layers, placed too close to other icons or text elements, or displayed against multiple backgrounds that reduce visibility).

Audio (single modality)

For deep fake audio-only content shorter than 30 seconds (e.g. commercials or advertisements), Signatories will include a short audible disclaimer at the latest at the beginning of the content.

For longer deep fake audio formats, real-time as well as non-real-time (e.g. audio-only social media content, AI-generated phone calls, AI-generated podcasts or radio broadcasts), Signatories will provide repeated audible disclaimers at the beginning, at appropriate intermediate phases, and at the end of the content.

Where deep fake audio content is delivered through a user interface and/or screen (e.g. car or smartphone display), Signatories will also display a visual icon or label via the elements of the user interface under their control, at the moment of the first exposure of the natural person, or upon initial access to the audio content (e.g. an icon or label embedded into a cover image or the title of the audio content).

Other Multimodal Content

For other multimodal deep fake content, Signatories will ensure that the multimodal content containing a deep fake is consistently disclosed using an icon or label, ensuring that the disclosure is clearly perceivable to the natural person without any further interaction on their part.

Other multimodal content includes, but is not limited to, the following combinations of static or dynamic content:

- image-text-audio;
- text-audio;

- image-audio;
- image-text.

Text

For AI-generated or manipulated text publications within the scope of Article 50(4) AI Act, Signatories will place the icon or label in a consistent position. This may be, for example, above or at the top of the text, near the headline of the text, or in the colophon at the beginning of the text, as long as placement is clear and distinguishable for the end-user within the type of text content published by the Signatory. If only part of the text publication is AI-generated or manipulated, it is sufficient to label only the part that is AI-generated or manipulated.

For short-form texts (single words or brief phrases), where labelling the text outputs would degrade readability, Signatories may ensure disclosure through contextual notice in the user interface or session (e.g., an indicator adjacent to the output, or session-level or page-level disclosure that AI was used).

Measure 1.3 Optional use of an EU icon and participation in its development

Recognising the fast pace at which technology develops, Signatories are encouraged to use the EU-wide icon (to be specified in the Annex) and to support the further development of an optional uniform EU label designed to provide more advanced and usable information on the AI-generated or manipulated elements of content. For this purpose, Signatories are further encouraged to support the work and activities of a dedicated task force to be facilitated by the AI Office aimed at advancing the development, usability and development of such a label, in accordance with the following requirements:

- The taskforce will be established after the publication of this Code of Practice with the participation of Signatories and relevant stakeholders from various sectors and fields of expertise, national competent authorities and existing Chairs/Vice-chairs of the relevant working group.
- The taskforce will assist and advise the AI Office in the further development and possible testing of the icon, including aspects related to accessibility and modality-specific solutions, such as an audio-only content.
- The taskforce's Signatories who make use of the icon will support usability and user feedback to ensure the icon remains clear and distinguishable for all natural persons, while following the design requirement set out in Measure 1.1 and allowing state-of-the-art technological advancements for future iterations.
- The icon will be freely available to all deployers and other users, AI system providers and online intermediary service providers. Technical elements of the EU icon will be provided under free and open-source licenses allowing distribution and use (such as the European Union Public Licence).
- The taskforce will support the refinement of the EU icon in a manner that avoids information overload and ensures that the disclosure remains meaningful and usable for all end-users.
- The task force will explore the possibility of developing an additional interactive element linked to the EU label (a "second layer") that will provide more detailed information about what has been generated or manipulated by AI. This second layer may be accessible by hovering over or clicking on the icon and may appear as a clear information field. Such interactive icon should follow the same design requirements as Measure 1.1, making it very

easy for end-users at all levels of literacy and digital skills to discern the information provided. Possible characteristics of such interactive icon are:

  - Disclose at the very top or beginning of the information field (second layer) that content has been AI-generated or manipulated and specify the type of manipulation (e.g., fully AI-generated content or specific modifications such as the removal of an object), through machine-readable marking techniques implemented in accordance with the Section 1 of this Code, where available, and/or through deployer-provided information;
  - The information in the second layer can be displayed in English and be available in all the languages of the Member States through a translation plugin so it can be displayed in the native language of the natural persons exposed to the content.

- To the extent technologically and practically feasible, the refined EU icon will be designed to work in tandem with and to further integrate the machine-readable marking and detection solutions as described in Section 1 of the Code. Its implementation will remain practical and proportionate across deployers of all sizes and operational contexts, while avoiding over-labelling.
- For audio-only disclosures, the taskforce will aim to test the accessibility and usability of disclosures (i.e. audio disclaimers, sound logos or signals), considering various vulnerable categories and the proportionality of disclosure time relative to total content duration.
- Further the task force could function as a forum for exchanging good practices of AI literacy for promoting the labels in all Member States with the support of the Signatories.

Commitment 2: Proportionate compliance, awareness and review

To ensure effective compliance with their obligations under Article 50(4) and (5) AI Act and the commitments and measures as specified in this Section of the Code, Signatories commit to implement proportionate internal processes, awareness measures and review mechanisms for the proper implementation of the labelling of deep fakes and text publications within the scope of Article 50(4) AI Act, taking into account their size and resources.

Measure 2.1: Internal compliance

Signatories will establish, adapt or maintain proportionate (existing) internal documentation or equivalent internal processes that specify how they implement the disclosure obligations under Article 50(4) and (5) AI Act. Such documentation or processes may include:

- A general description of the disclosures applied across services or products, in accordance with Commitment 1;
- A general description and representative, concrete and real examples of how disclosures are implemented in practice in accordance with Commitments 3 and 4, including deep fake content forming part of artistic, creative, satirical, fictional or analogous works. This description can clarify how the disclosure obligation under Article 50 (4) AI Act is applied to artistic, creative, satirical, fictional or analogous works, in accordance with Commitment 3. Or it can clarify when human review and editorial responsibility is involved in AI-generated or manipulated text publications on matters of public interest, in accordance with Commitment 4 and the related decision-making processes.

Deployers that regularly create deep fake content or AI-generated or manipulated text publications of public interest will also ensure appropriate oversight to review the proper application of the labelling obligations and the measures in this Section of the Code and to mitigate risks of non-labelled or incorrectly labelled content.

Signatories remain free to establish new processes or integrate the necessary processes into existing processes, including internal editorial and governance procedures, as required under applicable laws and professional standards, in accordance with the available organisational and technical capacities, provided that the disclosure obligations under Article 50(4) and (5) AI Act and the measures in this Section of the Code are fulfilled.

Measure 2.2: Awareness and Training

Signatories will make reasonable and proportionate efforts to ensure awareness of the disclosure obligations under Article 50(4) and (5) AI Act among personnel, including employees and external contractors, directly involved in the implementation of labelling measures or overseeing compliance with the measures in this Section of the Code.

Signatories are encouraged to provide training or equivalent guidance covering situations in which disclosure is legally required, how disclosures are implemented in the workflow, cases when editorial responsibility is involved or cases of artistic, creative, satirical, fictional or analogous work; accessibility considerations and procedures for correcting missing or incorrect labels when these have been identified.

Training should be proportionate to the size and resources of the Signatory and applied only to the extent to which personnel (considering their technical knowledge, experience, and education) are involved in creating, modifying, and disseminating relevant content.

Signatories remain free to determine the training formats and their frequency.

Measure 2.3: Review, feedback and cooperation with authorities

Signatories will support effective implementation of the disclosure obligations through review and feedback mechanisms.

Specifically, Signatories are encouraged to provide channels that allow individuals or third parties (trusted flaggers, independent fact-checkers etc.) to flag missing or incorrect disclosures, preferably through existing reporting mechanisms (e.g., trusted flagger mechanisms, interfaces for third-party fact-checking services or notice and action mechanism).

Signatories will review cases that have been reported or observed as mislabelled or non-correctly labelled and remedy disclosures without undue delay.

Signatories will cooperate with competent authorities in accordance with applicable European Union and national laws.

Commitment 3: Appropriate Disclosure for Artistic, Creative and similar Works

To fulfil their obligations in Article 50(4) and (5) AI Act, Signatories commit to implement measures to disclose deep fake content that forms part of evidently artistic, creative, satirical, fictional or analogous work or programmes.

Pursuant to Article 50(4) AI Act, such disclosure is 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, including its normal exploitation and use, while maintaining the utility and quality of the work. Where feasible, the disclosure should always travel with the content.

Signatories will use an icon, label or disclaimer following the design requirements in Measure 1.1. and will place it in a manner appropriate to the type of artistic, creative, satirical, fictional or analogous content and to the context in which it is presented. Such placement needs to be clear and distinguishable to end-users and provided at the latest at the time of first exposure to the content. Where relevant, this can be complemented with end credits, contextual or creative disclosure methods, and post-viewing disclaimers. The disclosure will be placed in a non-intrusive yet effective (i.e. clear and distinguishable) position, which may include, but is not limited to, the following:

- **Real-time or near real-time video**: at the latest at the time of the first exposure in the top or bottom corners for at least five seconds without further warnings throughout exposure (e.g. during opening credits);
- **Video**: the disclosure will be placed for a timing sufficient to inform the viewer at the latest at the time of first exposure without interfering with the experience (e.g. during opening credits);
- **Other multimodal content**: the disclosure will be displayed at the latest at the time of first exposure, ensuring that the disclosure is clearly visible to the natural person without requiring any further interaction on their part;
- **Image**: at the latest at the time of the first exposure in an appropriate place with the possibility of integrating it into the image or the background of the image while preserving the ability for the end-user to discern the labelling;
- **Audio**: an audible disclaimer should be inserted at the latest at the time of the first exposure.

Signatories can also conceive and implement ‘contextual’ disclosure solutions where the disclosure options mentioned above are not available or would affect the display or enjoyment of the work. When deep fake content is made available in a digital and/or interactive manner (e.g. on websites, apps or other user interfaces), the icon, label or disclaimer can be placed outside but adjacent to the video or image frame, or adjacent to the audio content and integrated into user interface elements or overlays under the control of the Signatories. Such a contextual disclosure solution should be perceivable by the end-user without the need for scrolling or additional engagement.

Where content is made available in a non-digital or non-interactive manner (e.g., exhibitions, art galleries, festivals or comparable contexts, audio or video on a physical carrier), disclosures can be provided, e.g. at the point of entry, when tickets are sold, or as part of introductory information or information provided via a physical carrier. Disclosure should be clear, accessible and understandable to all audiences. Where feasible, disclosure methods applicable to deep fake content made available digitally should remain attached to or travel with the content when it is shared or distributed.

Commitment 4: Human review, editorial control and responsibility in relation to AI-generated or manipulated text publications

To rely on the exception to the disclosure obligation in Article 50(4) subparagraph 2 AI Act, Signatories will establish, adapt or maintain minimal documentation, including existing procedures and documents, demonstrating that the AI-generated or manipulated text published for the purposes of informing the public on matters of public interest have undergone human review or editorial control prior to publication and that a natural or legal person holds editorial responsibility for the publication.

The procedures should be proportionate to the deployer's size and should include, at least, the following elements:

- identification of the natural or legal person with editorial responsibility (name, role and contact details);
- an overview of the concrete organisational measures as well as human resources allocated to ensure adequate human review or editorial control is performed and editorial responsibility is assumed before publication of the AI-generated and manipulated text publications.

Where not already publicly available, Signatories will publish the contact details of the natural or legal person with editorial responsibility to ensure accountability.

Signatories may optionally record additional information on the nature of the review or the type of involvement of the AI system in the generation or manipulation of the publication, where feasible.

To demonstrate compliance with this Commitment, Signatories may integrate internal arrangements into their existing review or editorial procedures to ensure alignment with their existing editorial quality checks and/or relevant professional standards, as applicable. Signatories who are media service providers subject to professional editorial standards and regulations may rely on those existing processes and responsibilities to demonstrate compliance with this Commitment.

The procedures and documentation adopted under this Commitment should in no way affect media freedom, editorial independence and protection of journalistic source information.

Appendices

Appendix 1 Sample Icon

This appendix contains suggestions for the icon or label with the acronym "AI" to be further discussed with participants.

Disclaimer: These sample icons only serve illustrative purposes and will be further developed throughout the drafting of the Code of Practice.

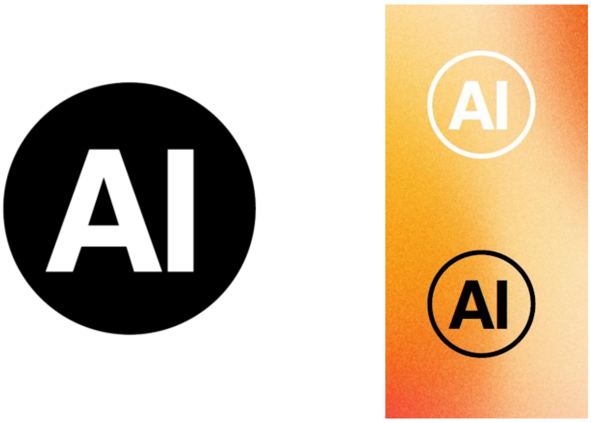

Figure 1. Sample icon developed by the European Commission (left) with variations against a background (right).

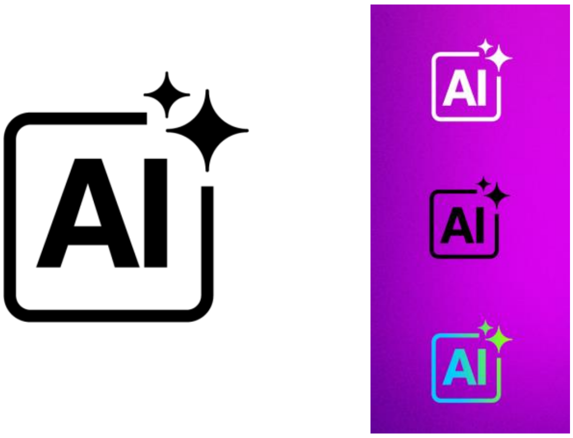

Figure 2. Sample icon developed by the European Commission (left) with variations against a background (right).