First 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 first draft of the Code of Practice on Transparency of AI-generated Content under the AI Act (the "Code"). This first draft of the Code addresses key considerations for providers and deployers of AI systems generating content falling 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)*

The version in your hands is the first draft of the Code which we present as a foundation for further refinement. Our work, which began in November 2025, has involved synthesising input from diverse stakeholders submitted in open public consultations and dedicated workshops. The Code of Practice process, and the resulting first draft, are unique: they are the result of a collaborative effort involving hundreds of participants from industry, academia, and civil society, as well as contributions from Member States. We have also been informed by the evolving literature on transparency of AI content, academic studies on the topic commissioned by the AI Office, relevant standards and international approaches, and the expertise and experience of Working Group members and observers.

Key features of the drafting process include:

- *A multi-stakeholder public consultation with 187 written submissions.*
- *Two specialised working groups led by us as Chairs and Vice-Chairs selected for our expertise, experience, independence and taking into account our geographical and gender diversity.*
- *Discussions and input received from the Code of Practice participants and observers through three workshops held on 17th and 18th November 2025 and relevant written input submitted in response to guiding questions.*
- *Review of relevant expert studies and other relevant documents on the topic.*

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 many of the relevant feedback points that we have received and to strike a balance between conflicting views.

Our task as Chairs and Vice Chairs of the working groups is to craft a Code that meaningfully integrates these insights while remaining true to the legal text and to the core purpose and obligations of the AI Act that the Code aims to facilitate:

- *to ensure that AI-generated and manipulated content are marked in a machine-readable and detectable manner with technical solutions that are effective, reliable, robust and interoperable, and*
- *to make it easier for natural persons to identify deepfakes and AI-generated or manipulated text which is published with the purpose of informing the public on matters of public interest.*

Although the first draft is not yet fully detailed, our approach aims to provide stakeholders with a clear sense of direction of the final Code's potential form and content, while we continue to engage in thorough deliberations regarding specific and concrete commitments and related measures.

At this stage, this first draft provides main commitments and measures, and remains high-level and broad. This is because i) we have focused initially on securing broad agreement on the structure, commitments and measures of the Code ii) there has been insufficient time to produce in this first draft detailed proposals with the level of consideration that such proposals would require for all issues at stake, and iii) we will update the draft Code's content to reflect latest developments on an ongoing basis and based on inputs from the participants in the process. In particular, this draft Code is expected to be further detailed and, where necessary, complemented or adjusted in future iterations.

***Following an iterative process of internal discussions within the Working Groups and additional input from stakeholders, Commitments and Measures may be added, removed, or modified in the future**. To provide even more insight into our deliberation, we have added open questions to highlight some of the areas where we aim to make progress in future drafts with the additional questions asked in the EUsurvey. This also serves the purpose of guiding feedback and submissions to allow various stakeholders to continue to participate effectively.*

Elements that need further development in the current draft, that we aim to address in the future versions, and for which specific stakeholder input is sought are as follows:

- *For WG1 (Section 1 of the draft Code):*
    - *Technical considerations on feasible approaches to marking AI-generated software code (as a specific type of AI-generated text).*
    - *Technical considerations on feasible approaches to marking other challenging kinds of content (e.g. very short texts) since marking them would reduce their quality and/or utility very significantly or setting thresholds to account for these limitations.*

    - *Technical considerations on the applicability of the currently proposed measures or the need to define new measures for agentic AI, gaming, VR, voice assistants, and other more novel kinds of AI-generated content.*
    - *Technical considerations on the implementation of Measures 2.1. and 3.4 in Section 1, and specifically on the creation of shared aggregated verifier(s).*
    - *Feedback on the clarity of the terminological definitions in the Glossary of Section 1.*

- *For WG2 (Section 2 of the draft Code):*
    - *Feedback on a common icon and the audio-only dimension thereof for the benefit of easy-recognition for the natural person both as an interim and long-term interactive solution. Such inputs could both be specific already implemented best practice icons as well as academic studies of citizens' discernment of AI content (covered by Article 50 (4) AI Act) with or without labels, or the use of similar labelling practices such as advertisement-labels including eye-tracking studies, experiments, surveys, focus groups or other qualitative studies.*
    - *Further input is needed, especially with regards to the technical solutions for the audio-only labelling, for the interactive function on what exactly has been AI-generated or manipulated, and the flagging system.*
    - *Even though we have tried to include the stakeholder input already received in the measures for the different modalities, these still need to be further refined. Input is sought not only on a general level, but also regarding specific decisions of icon placement and for specific modalities. This also goes for creative works.*

In formulating this first draft, we have been principally guided by the provisions in the AI Act for determining matters 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. Issues related to the scope of key definitions and exceptions from the obligations have not been addressed in this draft Code since they will 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 the current 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 second version of the Code, which will play a crucial role in facilitating transparency of AI-generated content in the EU. We welcome written feedback by the Code of Practice Plenary participants and observers ***by 23 January 2026 (22:00 CET), through the submission of feedback through the EUSurvey available to the Code participants.***

We are very much looking forward to the next stakeholder meetings in January and to the input that we will receive. They will help us ensure the Code remains true to the overarching purpose of the relevant provisions in the AI Act, namely to make it easier and effective for natural persons to identify and discern AI-generated and manipulated content.

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”), 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.

To achieve this overarching objective, the specific objectives of this 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 ensure providers of AI systems generating synthetic audio, image, video or text content comply with their obligations under Article 50(2) and (5) the AI Act and to enable the competent market surveillance authorities to assess compliance of providers of these AI systems 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 becomes increasingly hard for humans to distinguish AI-generated content from human-authored authentic content, impacting the integrity 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 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 legal requirements under Article 50(2) AI Act for effectiveness, interoperability, robustness and reliability. This calls for an implementation of a multi-layered approach, as this has been widely recognised as the best way of achieving a reasonable balance between the four legal requirements, as well as emerging as the most recommended marking approach based on the public consultation and expert stakeholder inputs. Therefore, an appropriate combination of marking techniques is to be applied to meet the requirements of Article 50(2) AI Act, as far as this is technically feasible for 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.

c) **Cooperation along the value chain:** Signatories recognise the need for practical arrangements for making, as appropriate, the detection mechanisms accessible and facilitating cooperation with other actors along the value chain, disseminating content or checking its authenticity and provenance to enable the public to effectively distinguish AI-generated and manipulated content. Signatories that are also generative AI model providers recognise the important role they occupy in the value chain to facilitate compliance by 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. The Signatories recognise that this Section should encourage providers of generative AI systems and models to advance the state of the art in AI marking and detection techniques and related processes and measures. The Signatories further recognise that if providers of generative AI systems or models can demonstrate equal or superior marking and detection techniques in compliance with Article 50(2) AI Act through alternative means that achieve greater efficiency, such innovations should be possible and recognised as advancing the state of the art in AI marking and detection.

e) **Cooperation with other stakeholders:** The 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 reliance on open standards and marking techniques implemented at the model level or provided by other third parties. The 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 the technical solutions. The 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. The Signatories further recognise the importance of cooperating with market surveillance authorities and of fostering collaboration between providers of generative AI systems and models, researchers, civil society and regulatory bodies to address emerging challenges and opportunities in the 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 content provenance marking standards need to be elaborated further to capture the provenance chain of content authoring, recording each creation or modification step carried out by an AI system.

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 and proportionate.

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) of the AI Act to mark in a machine-readable manner the outputs of generative AI systems, including general-purpose AI systems, Signatories commit to implement a multi-layered approach of active marking techniques with regard to the text, image, video or audio 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.

Signatories commit to implement such multi-layered approach as far as under the state of the art no single marking approach is sufficient to effectively comply with the requirements in Article 50 AI Act and to the extent this is applicable to the respective modality and type of content generated or manipulated by their AI system(s).

In order to fulfil this Commitment, Signatories commit to implementing the following Measures.

Measure 1.1: Machine-readable marking techniques

Signatories will implement marking techniques to ensure the outputs of their generative AI systems are marked with multiple layers of machine-readable marking as specified in this Commitment. For additional robustness, machine-readable marks may reference information that comes from different layers, e.g., a watermark may refer to a metadata identifier or vice versa.

The marking techniques can be implemented at different stages of the value chain (e.g., model providers) and can also be provided by third parties (e.g., providers specialized on transparency marking techniques). Signatories may rely on those technical solutions so long as the marking techniques are compliant with the requirements in this Section of the Code.

Sub-measure 1.1.1: Marking techniques for content that permits metadata embedding

If content is generated or exported in a data format that supports adding information as part of the metadata (e.g. an image, video, or document file), Signatories will add information about the provenance of the content and a signature of the generative AI system to the metadata. The metadata embedding will provide information about the type of the operation performed by the AI system (e.g., prompting, editing, or generation). The information will be digitally signed.

Sub-measure 1.1.2: Marking techniques interwoven within the content

Signatories will ensure that AI-generated or manipulated content is marked with an imperceptible watermark. This watermark will be directly interwoven within the content in a manner that is difficult for it to be separated from the content, and that withstands typical processing steps that may be applied to the content. Signatories will implement the watermark in the best possible technical and economically viable way.

Signatories may embed watermarks during model training, model inference, or within the output of an AI model or system. Signatories who provide AI models to other providers of AI systems will implement relevant marking techniques at the model level to facilitate compliance of downstream providers.

Sub-measure 1.1.3: Fingerprinting or logging facilities (where necessary)

Where necessary to address deficiencies in the marking techniques in Measures 1.1.1 and 1.1.2, Signatories will establish and maintain fingerprinting of the AI-generated or manipulated content or logging facilities 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, whereas for visual content perceptual hashing or other fingerprinting approaches may be preferable.

Measure 1.2: Marking techniques for specific modalities

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

Signatories will implement a digitally signed manifest allowing deployers to obtain a certified version of the AI output generated or manipulated by their AI system or model to formally guarantee the origin of content that does not allow secure embedding of metadata. This provenance certificate will enable deployers and end-users to provide third parties with guarantees that the content is AI-generated or manipulated, linking it back to the specific generative AI system or model.

Sub-measure 1.2.2: Marking of multimodal content

Signatories will ensure that multimodal output of their AI system is marked as specified in Measure 1.1. In addition, they will ensure that the employed marking techniques are synchronised across the modalities in a manner that the marking is recognisable when only one or a subset of modalities have been altered or exchanged.

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

Signatories releasing open-weight AI models or systems will implement structural marking techniques encoded in the weights during model training. This will facilitate third parties who use these open-weight models or systems to build generative AI systems to comply with Measure 2.2.

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

In order to facilitate compliance by downstream providers of generative AI systems, Signatories that are also providers of generative AI models will implement machine-readable marking techniques for the content generated or manipulated by their models prior to the model’s placement on the market.

To minimise cost and facilitate compliance, it is recommended that Signatories, including SMEs or SMCs that are providers of AI systems, use one or more generative AI models which already mark the outputs in a manner compliant with the relevant measures in Section 1 of the Code. However, it is the responsibility of the Signatory to ensure that all AI-generated or manipulated outputs are suitably and compliantly marked, especially in the case of multimodal outputs or when outputs of multiple generative AI models are combined.

Measure 1.5: Non-removal of machine-readable marking

Signatories will implement appropriate measures to preserve marks and other intrinsic provenance signals on AI-generated or manipulated content by:

a) ensuring that existing detectable marks are retained and not altered or removed, including where such content is used as input and subsequently transformed by their AI system into a new output, and
b) include in the acceptable use policy, terms and conditions or the documentation accompanying their generative AI system or model a prohibition for removal or tampering of the marks by deployers or any other third parties.

Measure 1.6: Transparency of the provenance chain

Signatories will record and embed through content marking the origin and provenance chain from AI-assisted or (partially) modified content to fully AI-generated content where technologically possible for the specific modality. To this end, Signatories will consistently check the marking and provenance information of the inputs to their AI systems. They will add or record all content provenance steps within their AI systems (both AI and human) in a way that distinguishes the additional operation from previous operations, by leveraging metadata or other appropriate techniques, where technically feasible to record and verify the provenance chain.

Signatories are also encouraged to record provenance information for fully human-authored content or fully human content editing operations in order to increase trust and facilitate authenticity and provenance of all content.

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

In order to facilitate compliance by deployers of generative AI systems with their obligations for distinguishable disclosure of deep fakes and certain AI-generated and manipulated text under Article 50(4) AI Act, Signatories who provide generative AI systems will provide a functionality in their system's interface and implement an integrated option that allows deployers to directly – upon generation of the output – include a perceptible mark or label in the content enabled by default.

Signatories will implement such perceptible marks and/or labels in consistency with the Commitments and Measures in Section 2 of the Code.

Signatories will also implement supporting measures for display of labels and provenance metadata that enable deployers, platforms and websites to implement display practices and policies that are appropriate for their use cases.

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

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 obligation under Article 50(2) AI Act to ensure that the outputs of their AI system(s) are detectable as AI-generated or manipulated, Signatories commit to implementing the following measures to enable the detection of text, image, video or audio content, or a combination thereof, as generated or manipulated by their AI system or model.

Measure 2.1: Enable detection by users and other third parties

Signatories will provide free of charge an interface (e.g. API or user interface) or a publicly available detector to enable users and other interested parties to verify with confidence scores whether content has been generated or manipulated by their AI system or model. For marking and detection techniques that provide information about the provenance from other AI system providers (e.g., in the metadata), the interface will also disclose a complete set of the provenance information.

Signatories are encouraged to collaborate with the Commission and other relevant actors to make the detection mechanism directly available in distribution and communication platforms and to maintain these mechanisms throughout the system and model's lifecycle.

In case a Signatory goes out of business, they will make the detectors 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: Detectors for already marked AI-generated content produced by a generative AI model

In order to facilitate compliance by downstream providers of generative AI systems, Signatories who are also providers of generative AI models will provide detection mechanisms for the content generated or manipulated by their models prior to the model's placement on the market.

Measure 2.3 Forensic detection mechanisms

To complement the marking techniques specified in Commitment 1 and as an additional line of defence and alternative to Measure 1.1.3., Signatories who are providers of generative AI models that can be used or integrated into downstream generative AI systems will implement forensic detection mechanisms which do not depend on the presence of active AI marking.

Signatories are also encouraged to collaborate with competent market surveillance authorities and, as appropriate, research organisations and other relevant stakeholders, to support the development of an aggregated forensic detector that is capable of detecting the outputs of generative AI models or integrated into AI systems available on the Union market.

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

Signatories will embed in the results of their marking and detection solution human-understandable explanations of the evidence for the detection and provenance results, as far as technically feasible.

Where applicable, 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. 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".

Measure 2.5: Support literacy for AI content provenance and verification

Signatories will provide documentation, training materials, and other relevant information to support deployers and other users in making informed decisions on what marking and verification tools they may use, including helping them to understand how to access and apply detection mechanisms and to interpret the provenance data and the detection results.

Signatories are encouraged to collaborate with academia, civil society and other relevant organisations to promote literacy and awareness regarding AI content provenance and verification.

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.

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

LEGAL TEXT: Article 50(2) and recitals 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 detection of AI-generated or manipulated content are effective, robust, reliable and interoperable, 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 comply with these requirements in a balanced manner and ensuring that all requirements are met, as outlined in the following measures.

Signatories commit to implement the measures and meet the requirements prior to placing their generative AI system or model on the market or putting it into service, and throughout their lifecycle.

Measure 3.1: Effectiveness

Signatories will implement technical marking and detection solutions that are fit-for-purpose and capable of effectively informing about the artificial origin of the content, and that contribute to the integrity of the information ecosystem. To this end, Signatories will implement marking and detection solutions that are computationally efficient and low-cost, that ensure real-time application, and that are capable of preserving the quality of the generated content, without compromising the functioning of the AI models or systems and while aiming for environmental sustainability.

Measure 3.2: Reliability

Signatories will implement marking and detection solutions that are reliable and aligned to the state of the art. In particular, for what concerns the accuracy of the detection of AI-generated or manipulated content, reliability will be measured using relevant established metrics, such as false positive /negative rates of the detection and bit error rates in decoded marker information (if applicable). Low false-positive and false-negative rates need to be demonstrated on samples of AI-generated and human-authored content unseen during the training and development of the AI models or systems.

Measure 3.3: Robustness

Signatories will implement marking and detection solutions that achieve a high level of robustness of the marking technique to common alterations and adversarial attacks. Common alterations include typical processing operations such as mirroring, cropping, compression, screen capturing, paraphrasing, character deletions, changes in image or video resolution, pitch shifting, time stretching, or change of format. Signatories will assess security robustness in terms of resilience to adversarial attacks such as copying, removal, regeneration, modification, and amortisation attacks on the markings. When providing access to the detector or the detection interface, Signatories will apply standard security practices such as rate limits, to prevent and counteract malicious use and attacks against the marking and detection mechanisms.

Measure 3.4: Interoperability

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

Signatories who develop their own marking and detection solutions will collaborate towards the creation of shared aggregated verifier(s) to detect outputs of their generative AI systems or models. Alternatively, they will implement other appropriate measures to directly encode in the output or its metadata information about the means that can be used to detect and verify the machine-readable marks, to ensure that the detection methods are interoperable across providers and easily accessible for users.

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 to promote interoperability and broad adoption, and to minimise costs of compliance.

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.

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

Contingent upon their capacity and resources, Signatories will invest in scientific research and collaborate with 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.

Specifically, Signatories are encouraged to cooperate on the development of watermark schemes that enable controlled renewal, revocation, or replacement without degrading the quality of the original output, and the development of future forensic models and fingerprinting techniques.

Commitment 4: Testing, verification and compliance

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

In order to effectively fulfil and demonstrate compliance with their obligation under Article 50(2) and (5) and the commitments and measures as specified in this Section of the Code, Signatories will set up, keep up to date and implement testing, verification and compliance framework, 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 the Commitments and Measures in this Section.

The framework will contain a high-level description of implemented and planned processes and measures to adhere to this Section of the Code and 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.

When 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 assume responsibility and commit to demonstrate compliance with this Section of the Code and Article 50(2) and (5) of the 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, including, where appropriate, by involving independent experts and/or in the context of AI regulatory sandboxes under regulatory supervision.

In the context of testing and evaluation, Signatories will take into account available 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.

To ensure that the 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 and update detectors to keep false positive rates low, while preserving detectability.

Signatories will implement and document appropriate follow-up actions on reported instances of malfunctioning, adversarial attacks and compliance shortcomings by deployers, independent researchers and other third parties.

Measure 4.3: Training

Signatories will provide appropriate training to personnel involved in the design and development of AI systems and models and overseeing the compliance to ensure 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 to demonstrate compliance with their commitments under the Code and provide all relevant information and access to the system.

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 metadata entry. The purpose of this addition is to facilitate detection of this marking and provenance attribution of the AI-generated or manipulated content.
Active detection	Verification of markings such as watermarks or metadata markers that have been purposefully added by a provider of an AI system or model.
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 verification of authenticity of the provider and integrity of the signed content.
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 authentic ones.
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.
Open-weight Model	A model where the underlying weights, code, and parameters are made publicly available.
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 steps 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.
Structural marking	An imperceptible watermark that is either embedded into a model during training or upon inference. This can be a technique to add a marking to an open-source model that can be downloaded from the internet. However, in this case its security is inherently limited because the watermarking key must at least implicitly be shipped with the model.
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 AI system of a provider or a person exposed to the content.
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 of deepfakes and AI-generated and manipulated 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 of deepfakes and AI-generated and manipulated 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.

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 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 deployers of an AI system that generates or manipulates 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 and manipulative 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 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) **Context of dissemination:** Signatories acknowledge that as deployers of AI systems generating or manipulating content, they are responsible for labelling the output accordingly and for disclosing its artificial origin or manipulation in a manner that is appropriate to the context of dissemination. These responsibilities are additional and 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 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 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 users, particularly in relation to different needs and vulnerabilities. Such icons 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 content is essential for individual awareness and for supporting AI literacy. Public awareness and transparency about AI-generated or manipulated content and detection tools 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 (DSA), intellectual property, media (AVMSD and European Media Freedom Act), political advertising and other relevant regulatory frameworks.

Part A: General Commitments

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.

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

In order to fulfil their obligations in Article 50(4) AI Act, Signatories who are deployers of AI systems that generate deep fakes or text publications falling within the scope of Article 50(4) and Article 2(1) AI Act commit to apply consistent disclosure of origin and to use a common taxonomy and icon as specified in the following measures.

Measure 1.1: Implement a common taxonomy

Signatories will use a common taxonomy to support consistent and transparent identification of content falling within Article 50(4) AI Act that provides a harmonised vocabulary and serves as a guidance tool to distinguish content that triggers disclosure obligations under Article 50(4) of the AI Act.

To signal the granularity of the deceitful elements for the natural persons exposed to the content, the Signatories will use a common taxonomy for classifying content that qualifies as a deepfake or as AI-generated or manipulated text publications under Article 50(4), particularly distinguishing between the following categories of content**:**

Fully AI-generated content

covers content fully and autonomously generated by the AI system without human authored authentic content (e.g., fully AI-generated images, video or audio based on prompts to the system; AI-generated books, articles or other content on matters of public interest, including political or social issue texts intended to persuade).

AI-assisted content

covers content with mixed human and AI involvement, where AI-assisted image, audio or video generation or modification affects meaning, factual accuracy, emotional tone, or other elements that may falsely appear authentic or truthful; or as regards AI-generated or manipulated text published on matters of public interest, where the AI system substantially impacts the content of text publication. Generating or altering the content for the purpose of this Measure could include, but is not limited to:

- object removal;
- face/voice replacement or modification;

- adding AI-generated or manipulated text to existing human-authored text or onto real images;
- hybrid audio formats combining deep fake audio and authentic audio, significant visual or audio alterations, including beauty filters that change perceived age and/or emotional tone;
- AI rewriting or summarising human-created text;
- AI-generated text that imitates the style of a specific person;
- AI-enabled alterations to description of events, actors, arguments or interpretation;
- seemingly small AI-alterations that change the context of the content (e.g. noise removal that makes it appear as though the interviewee is in a different setting), editing that changes background information; or colour adjustments that change contextual meaning (e.g. skin tone).

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

The Signatories will apply a common icon for deepfakes and AI-generated and manipulated text publications as a method of disclosure. They will place it in a visible and consistent location appropriate to the context.

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

Until an EU-wide icon is finalised, the Signatories may use an interim icon to support consistent disclosure composed of a two-letter acronym referring to artificial intelligence, which can also be letters referring to the translation into the languages of the Member States (e.g. AI, KI, IA), as illustrated by the sample icons contained in Appendix 1.

The icon should be:

- Clearly visible at the time of the first exposure.
- Placed in a position appropriate to the content format and dissemination context.
- Implemented in a way that does not interfere with the enjoyment of artistic, creative, satirical or fictional works.
- Include the two-level taxonomy as defined in Measure 1.1.

Further practical details regarding the use and placing of the icon per content modality can be found in Sections B and C.

Sub-measure 1.2.2: EU common icon

Signatories commit to support the development of a common interactive EU icon that will be further explored following usability tests and interoperability requirements. The icon should:

- be designed in a way that integrates the possibility for the natural person to distinguish different degrees of deceitful content as specified in the taxonomy from Measure 1.1;
- make it possible, when interacting with it, to get further information of what exactly has been AI-generated or manipulated, in accordance with the machine-readable information provided by the marking as specified by Article 50(2) and Section 1 of this Code;
- support consistent disclosure composed of a two-letter acronym referring to artificial intelligence, which can also be letters referring to the translation into the languages of the Member States (e.g. AI, KI);

- be placed in a fixed position appropriate to the content format and dissemination context;
- be implemented in a way that does not interfere with the enjoyment of artistic, creative, satirical or fictional works;
- potentially include interactive audio disclosures that could be integrated in audio-only content and for accessibility purposes, and comply with applicable accessibility requirements.

Signatories will support the process of developing the common interactive EU icon facilitated at the EU level.

Commitment 2: Compliance, training and monitoring

To effectively fulfil and demonstrate compliance with their obligations under Article 50(4) and the commitments and measures as specified in this Section of the Code, Signatories will set up, implement, keep up to date a compliance and monitoring documentation (proportionate to the size and resources of the deployer) as well as cooperation mechanisms with competent authorities, as specified in the following measures.

Measure 2.1: Internal compliance

Signatories commit to draw up, keep up to date and implement internal compliance documentation that specifies their labelling practice i.e. how they have applied labelling requirements, including the icon in Measure 1.2, and provide concrete examples of how they have used it.

This compliance process will integrate the specific compliance measures specified in Parts B and C of this Section of the Code regarding deepfakes and AI-generated and manipulated text publications.

Measure 2.2: Training

The Signatories will provide appropriate training to personnel involved in the creation, modification, or distribution of content covered by Article 50(4) AI Act. Training should cover at least:

- when disclosures are legally required;
- how to apply the taxonomy and icon (when used);
- understanding the specific disclosure practices for artistic and creative work and for the exception for text publications the application of the requirement for human review and editorial responsibility;
- applicable accessibility requirements for disclosures; and
- procedures for correcting missing or incorrect disclosure.

Training should be proportionate to the Signatory's size, to the resources of the Signatory and to the risks associated with the content generated or manipulated by the AI systems used, taking into account its context, the extent of its dissemination and its potential impact.

Measure 2.3: Monitoring and cooperation with market surveillance authorities

Signatories will facilitate the possibility for third party and natural persons to, confidentially and in a secure channel, flag mis-labelled or non-labelled deepfakes and AI-generated and manipulated text of public interest.

Specifically, the Signatories commit to duly cooperate with market surveillance authorities and other third parties that have an interest in understanding and/or assessing whether content was duly labelled. Other third parties include media regulators, providers of intermediary services, including Very Large Online Platforms and Very Large Online Search Engines as defined in the Digital Services Act, certified independent fact-checking organisations (e.g. members of EFCSN and IFCN).

The solution should allow for Signatories to document that they labelled consistently and timely and follow up on reported instances of non-compliance, including to facilitate the reporting of unlabelled or mislabelled deepfakes and AI-generated and manipulated text publications across member states.

Once flagged and assessed as mislabelled or incorrectly non-labelled, the Signatory will fix missing or incorrect icons or labels without undue delay.

Commitment 3: Ensure Accessible Disclosure for all Natural Persons

The Signatories commit to ensure icons with associated labels are accessible and conform to applicable accessibility requirements under Union law. Signatories commit to comply with or facilitate compliance with applicable accessibility requirements by themselves or, respectively, actors whose products or services enter into the scope of such requirements.

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

Where necessary, Signatories will actively support relevant actors involved in the creation, distribution, or oversight of content with appropriate technical and organisational measures, following state-of-the-art accessibility protocols, procedures and technological standards. In particular, this may require visual icons to comply with contrast, size and screen-reader standards, and audio cues to be provided for visually impaired users. Signatories will provide:

- audio descriptions for visual indicators;
- visual/ tactile cues for audio-content only;
- high contrast icons and screen-reader compatibility.

Where relevant, Signatories will ensure labels are accessible to screen-readers and available in alternative modalities (e.g., alt-text, metadata).

Signatories will conduct a self-assessment of the applicable accessibility requirements or will use an EU-wide or other standardised icon which has been assessed for its conformity with those requirements.

To comply with this measure, Signatories are encouraged to provide support 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”.

Part B: Specific Commitment and Measures for Deepfakes

LEGAL TEXT: Article 50(4) and 50(5) and recitals 134 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.

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 specific commitment and measures for 'deep fakes' specified in this Part B apply in addition to the general commitments and measures specified in Part A of this Section of the Code.

Commitment 4: Specific Measures for Deepfake Disclosure

To fulfil their obligation in Article 50(4) and (5) AI Act regarding to deep fakes, Signatories commit to implement the following measures to correctly classify 'deep fake' content and ensure clear, distinguishable and timely disclosure.

Measure 4.1: Internal processes for consistent classification of Deepfake Content

Signatories will set up and implement internal processes to identify deepfake image, audio, video content and apply the definition of "deepfake" in Article 3(60) in a consistent manner and to determine whether applicable exception apply (e.g. law enforcement use) or if the content relates to artistic, creative, satirical and fictional work.

They will take into account the target audience of the content, the specificities of the distribution channels they employ and any other relevant elements.

Signatories will ensure that the deepfake labelling process is not only based on automation but also supported by appropriate human oversight.

Measure 4.2: Clear and distinguishable disclosure of deepfakes

In accordance with their obligation in Article 50(5) AI Act, Signatories will disclose the deep fake content in a clear and distinguishable manner at the latest at the time of the first exposure, by applying the following sub-measures for different modalities of content and multimodal content.

Sub-measure 4.2.1: Real-Time Video

For real-time deepfake video, Signatories will display the icon in a non-intrusive way consistently throughout the exposure where feasible.

Furthermore, Signatories will insert a disclaimer at the beginning of exposure that explain that this display content includes deepfake. This disclaimer should be perceivable for an appropriate duration.

Sub-measure 4.2.2: Non-Real-Time Video

For non-real-time deepfake video, Signatories will disclose that the video contain deepfakes with the icon. The Signatories may choose among the following disclosure options, individually or combined, as appropriate to the context.

- A disclaimer at the beginning of the exposure. In case of an oral disclaimer, the icon needs to appear simultaneously with the audio disclosure.
- Placing the icon consistently throughout the exposure in an appropriate fixed place, ensuring that the disclosure is clearly visible to the natural person without any additional interaction. For online platforms, this entails placing the icon consistently just outside the video frame integrated into the user interface or interface overlay.
- A disclaimer in the credits at the end of the video. This measure always needs to be accompanied by one of the three previous measures.

Sub-measure 4.2.3: Other Multimodal Content

For other multimodal deepfake content, Signatories will ensure that the multimodal content that contains deepfake is consistently disclosed using the icon, ensuring that the disclosure is clearly visible to the natural person without any further interaction on their part.

Other multimodal content includes, but is not limited to, the following combinations ofstatic or dynamic content, excluding content covered by Measures 4.2.1 and 4.2.2:

- image-text-sound;
- text-sound;
- image-sound;
- image-text.

Sub-measure 4.2.4: Image (single modality)

For deepfake images, Signatories will place the common icon consistently at any exposure in a fixed place. The icon should be clearly distinguishable and visible, particularly from the image itself, and not be hidden, as in the case of image layers or multiple backgrounds.

Sub-measure 4.2.5: Audio (single modality)

For deepfake audio-only content shorter than 30 seconds (e.g. commercials/ads), Signatories will include a short audible disclaimer, in plain and simple natural language, of the content disclosing the artificial origin of the deep fake audio at the beginning of the content.

For longer audio formats such as podcasts, Signatories will provide repeated audible disclaimers at the beginning and intermediate phases, and at the end of the content.

When a screen is available in the audio-interaction with the user (e.g., car or smartphone display), Signatories will also display the icon at the moment of first exposure of the natural person or upon initial access to the audio content.

In case the EU-wide icon (once available) includes an audio-only solution, this solution could be used instead of the natural language disclaimer.

Measure 4.3: Apply Appropriate Disclosure for Creative Works

With regard to deep fake content that forms part of evidently artistic, creative, satirical, fictional or analogous work or programmes, Signatories will disclose such deepfakes in an appropriate manner that does not hamper the display or enjoyment of the work, including its normal exploitation and use, while maintaining the utility and quality of the work and appropriate safeguards for the rights and freedoms of third parties.

Signatories will place the icon from the Commitment 2 in a non-intrusive position. Non-intrusive positions include but are 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 or for at least five seconds without further warnings throughout exposure;
- Video: the icon should be placed for a timing sufficient to inform the viewer at first exposure without significantly interfering with the experience;
- Other multimodal content: the icon 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 user to discern the icon;
- Audio: When a screen is available, placing the icon in an appropriate place, at the latest at the beginning of exposure, will suffice, without requiring an audible disclaimer at the beginning of the content. When no screen is available, a non-intrusive audible disclaimer should be inserted at the latest at the time of the first exposure. Such non-intrusive audible disclaimers may include, but are not limited to, a potential EU-wide audible disclaimer using either spoken disclosure (in the same language as the content), rhythmic cues, or sound-based signals.

When content forms part of evidently artistic, creative, satirical, fictional or analogous work or programme, Signatories will apply appropriate safeguards for the rights and freedoms of third parties. Third parties include depicted or simulated persons and a wider audience if the deepfake addresses political or societal sensitive topics. Safeguards include but are not limited to disclaimers of the deceitful element and applicable requirements under other Union and Member States law, in particular to avoid violations of the depicted or simulated persons' privacy, dignity and other fundamental rights and freedoms.

Part C: Specific Commitment and Measures for AI-Generated and Manipulated Text

LEGAL TEXT: Article 50(4) and 50(5) and recitals 134 and 135 AI Act

4. 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 specific commitment and measures for AI-generated and manipulated text in this Part C apply in addition to the general commitments and measures specified in Part A of this Section of the Code.

Commitment 5: Specific Measures for Disclosure of AI-Generated or Manipulated Text

In order to fulfil their obligation in Article 50(4) and (5) AI Act with regard to AI-generated or manipulated text, Signatories commit to implement the following measures in order to correctly identify all AI-generated or manipulated text published with the purpose of informing the public on matters of public interest, where no human has reviewed the text publication and no natural or legal person has assumed editorial responsibility (hereafter 'AI-generated and manipulated text publications') and to ensure clear, distinguishable and timely disclosure.

Measure 5.1: Internal Processes For AI-Generated and Manipulated Text

Signatories will set up and implement internal processes to correctly identify AI-generated or manipulated text publications in a consistent manner, taking into account the target audience of the content and the specificities of the distribution channels they employ, and to determine whether legal exceptions apply (e.g., law-enforcement use or human review/editorial control with editorial responsibility).

Signatories will ensure that the AI-generated and manipulated text publication labelling process is not only based on automation but also supported by appropriate human oversight.

Measure 5.2: Clear and Distinguishable Disclosure

In accordance with their obligation in Article 50(5) AI Act, Signatories will disclose the AI-generated and manipulated text publications in a fixed, clear and distinguishable manner at the latest at the time of the first exposure:

Signatories commit to place the icon in a fixed, clear and distinguishable position. This fixed place could include but is not limited to placing the icon at the top of the text, beside the text, in the colophon or after the closing sentence of the text.

Measure 5.3: Human Review, Editorial Control and Responsibility

To rely on the exception in Article 50(4) subparagraph 2 of the AI Act and avoid disclosure of AI-assisted text publications, Signatories will establish internal procedures and maintain minimal documentation demonstrating that the AI-generated or manipulated text publications have undergone human review or editorial control and that a natural or legal person has editorial responsibility.

The procedures and documentation 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 and technical measures as well as human resources allocated to ensure adequate human review or editorial control is performed before publication of the AI-generated and manipulated text publications, including consideration of national specificities where relevant (e.g., linguistic, cultural or context-specific factors that may affect interpretation or impact);

- the date of the review and approval;
- a reference to the final approved version of the content (e.g., file name, URL, or any other internal identifier).

Signatories may optionally record additional information, such as the nature of the review or the type of AI involvement when feasible without creating an administrative burden.

Appendices

Appendix 1 Sample Icon

This appendix contains inspirations for the interim icon and the interactive EU-wide icon.

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

The indication of the two-level taxonomy in an icon with the AI-acronym

To support consistent disclosure, a two-letter acronym referring to artificial intelligence should be used, which can also reflect the relevant translation in the languages of the Member States (e.g. AI, KI, IA).

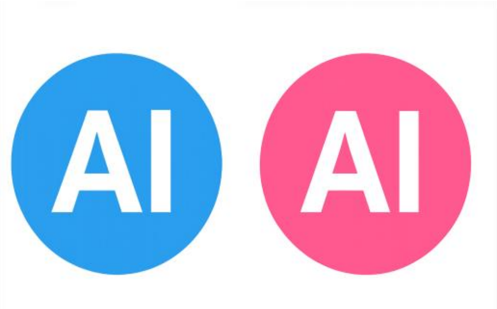


Figure 1. A zero-shot prompt on ChatGPT's free version as of December 2025 for an icon containing the word AI in two different colours indicating the difference between fully automated and AI-assisted content.

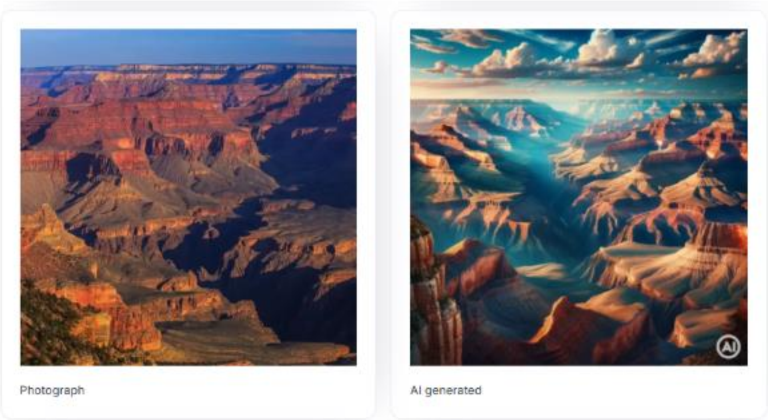


Figure 2. A round icon containing “AI” in the bottom right corner of the AI-generated photo. Source: Centre for AI Safety (CAIS)

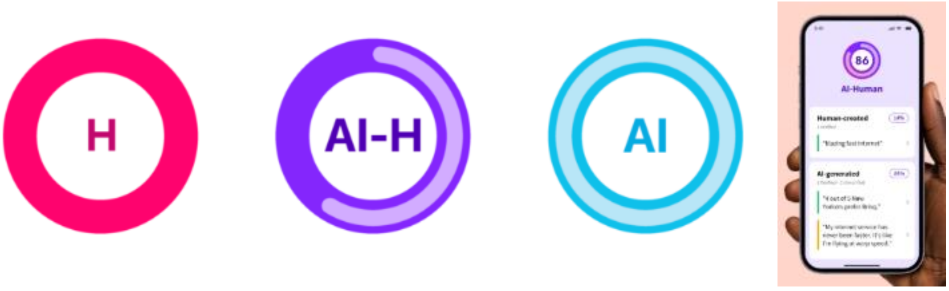
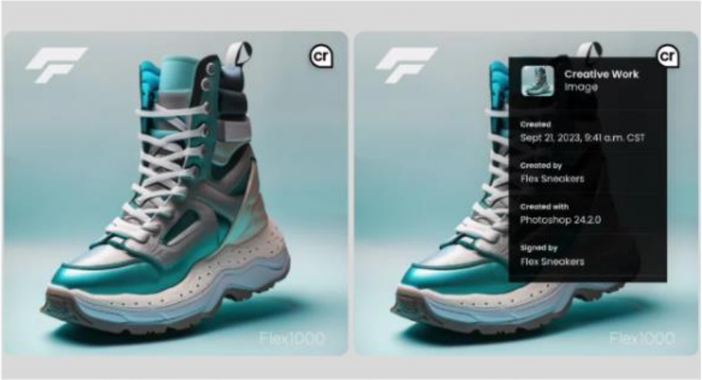


Figure 3. An icon has been developed by Artifact studio, indicating through colors and acronyms whether it is fully AI-generated or AI-assisted (here AI-H(uman) or human created (disregarded in the context of 50(4). Furthermore, the picture on the right shows interactive function with more information on what has been altered for the AI-H icon. Source: Fastcompany - https://www.fastcompany.com/90903238/simple-icon-it-easy-to-spot-ai-generated-content.

The interactive function when clicking or hoovering over the icon with the possibility of having explained what has been AI-generated or manipulated

For the AI-assisted content the interactive function will provide information on what has been AI-generated or manipulated using for instance machine readable information from article 50(2).



Creative Work
Image

Created
Sept 21, 2023, 9:41 a.m. CST

Created by
Flex Sneakers

Created with
Photoshop 24.2.0

Signed by
Flex Sneakers

Hovering over Adobe's CR icon will bring up general metadata of the image. Photo: Adobe.

Figure 4. Adobe has created an icon into which special metadata or “Content Credentials” can be embedded – this could also contain what has been AI-manipulated or generated.