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

Section 1: 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 1 Vice-Chair

Christian Riess
Working Group 1 Vice-Chair

Section 1: 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 ("generative AI systems") in complying with their obligations under Article 50(2) and (5) AI Act and

c) to enable the competent market surveillance authorities to assess such compliance in relation to providers who rely on the Code to demonstrate compliance with these obligations under Article 50(2) and (5) AI Act in a consistent, predictable and uniform manner across the Union.

Recitals

Whereas:

a) **Trust in the information ecosystem:** Signatories recognise that AI systems can generate and manipulate large quantities of synthetic content and that it is becoming increasingly difficult for humans to distinguish this content from human-authored content. Such developments impact the integrity of the information ecosystem as well as the trust people have in it, and raises new risks of misinformation, 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) **Technical solutions for marking:** While at the time of publication of this Code several techniques are readily available for marking AI-generated or manipulated audio, image, video, or text content, Signatories acknowledge that in most cases, in particular for content that can be disseminated online, no single marking technique suffices to meet the four requirements in Article 50(2) AI Act, namely effectiveness, interoperability, robustness, and reliability. In such cases, Signatories recognise that, as acknowledged in scientific reports and in line with the state of the art, only an appropriate combination of marking techniques

and associated detection mechanisms can allow satisfaction of those requirements in a holistic manner, in so far as this is technically feasible, taking into account the specificities and limitations of the content, the cost of implementation, and the existence of relevant standards. The Code promotes innovation and future-proofness in a fast-evolving technological landscape by allowing Signatories 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 results and common benchmarks, where available.

c) **Cooperation along the value chain:** Signatories recognise that effective, interoperable, robust, and reliable technical solutions for marking and detection require investment of time and resources. Signatories therefore acknowledge the need for practical arrangements to make detection solutions accessible, as appropriate, and for facilitating cooperation with other actors along the value chain to enable the public to effectively identify AI-generated or manipulated content. Signatories who are providers of generative AI models play an important role in this value chain and are therefore encouraged to facilitate compliance for downstream providers of generative AI systems built on those models.

d) **Advancing innovation in marking techniques and detection mechanisms:** Signatories acknowledge that determining the most effective, robust, reliable, and interoperable technical methods for marking and detection remains an evolving challenge. Signatories recognise that this Section of the Code promotes innovation in various ways and encourages advancing the state of the art in AI transparency techniques and related processes and measures. For instance, the Code allows (i) innovative techniques for AI transparency, (ii) new means to facilitate interoperability among different marking techniques and associated detection mechanisms, and (iii) the development of benchmarks to evaluate the performances of the technical solutions implemented by providers of generative AI systems. Signatories are therefore encouraged to invest time and resources, as appropriate, to contribute to the advancement of the state of the art.

e) **Cooperation with other stakeholders:** Signatories recognise the advantages of collaborative efficiency, for example through 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 end-users to detect marked content, and of engaging expert or lay representatives of civil society, academia, and other relevant stakeholders in understanding technical solutions for AI transparency. Signatories recognise that such cooperation may involve providing the necessary support for the smooth integration of third-party models and marking solutions into the AI system's inference process, and entering into agreements to share information relevant to technical solutions, while ensuring proportionate protection of sensitive information and compliance with applicable Union law. Signatories also acknowledge that end users' feedback on their implementation of marking and detection techniques is a valuable source of information in view of continuously improving its implementation. 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, search engines, researchers, civil society and regulatory bodies to address emerging challenges and opportunities in transparency of AI-generated or manipulated content.

f) **Promoting standardisation:** Signatories recognise the value of open standards to facilitate interoperability and lower the cost of marking techniques and associated detection mechanisms for AI transparency. 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. Signatories are encouraged to participate in relevant standardisation activities in order to accelerate the consolidation of marking techniques and associated detection mechanisms and their interoperability. In particular, Signatories recognise that metadata marking standards need to be further elaborated to provide richer information for AI transparency, as well as recognising the need for relevant watermarking standards both at technology and interface levels.

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 to Signatories who are providers of generative AI systems falling within Article 2 and Article 50(2) AI Act, taking into account the scope of the legal obligations and relevant exceptions, as applicable.

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 Union market independently from AI systems and third-party providers of marking and detection 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, with a view to facilitating compliance by downstream providers of AI systems built on those models and/or using third parties' marking and detection solutions. In such cases, the commitments and the measures of this Section of the Code shall be understood as applying to the Signatory's generative AI model and/or marking and detection solution as appropriate.

In this Section, the terms "will", "encouraged" and "may" should be interpreted in the following way:

1. **'will'** is used for mandatory measures under the Code that need to be met for the Signatory to be compliant with Article 50(2) and (5) AI Act, and for which compliance will be monitored by competent market surveillance authorities;
2. **'encouraged'** is used for optional measures under the Code that are not legally required and are purely voluntary, but are nevertheless recommended;
3. **'may'** is used for optional measures under the Code that are not legally required and are purely voluntary or provide providers with flexibility on how to implement the respective measure or commitment.

Commitment 1: Marking of AI-generated or Manipulated 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 a marking solution with regard to 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 type(s) of content generated or manipulated by their AI system(s).

Measure 1.1: Machine-readable marking techniques

Signatories will implement a marking solution that consists of at least one machine-readable marking technique, as detailed below, which, in conjunction with their respective detection mechanism, meets the level of effectiveness, reliability, robustness, and interoperability required by Article 50(2) AI Act and Commitment 3 of the Code.

So long as no single marking technique can, under the state of the art, ensure by itself compliance with the four requirements in Article 50(2) AI Act of effectiveness, interoperability, robustness, and reliability for audio, images, video, and containerised text, in particular for content that can be disseminated online, 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 marking, as specified in Sub-measures 1.1.1 and 1.1.2 below.

To enable a proportionate implementation of the above requirements, a single layer of marking will be considered sufficient in specific cases where a generative AI system is embedded in physical products capable of generating synthetic outputs in a technically controlled and closed environment mainly instructive in nature. This applies to the extent effective technical measures are embedded in the product to prevent the output from leaving the environment of the product, for example by being captured or exported and disseminated externally. Moreover, given that free-form text cannot transport metadata, a single-layer of marking as described in Sub-measure 1.1.2 is considered sufficient to comply with the requirements of Article 50(2) AI Act for this specific type of content.

In the future, Signatories will also be able to demonstrate compliance with Article 50(2) AI Act through the use of alternative marking technique(s) (possibly a single marking technique), provided that they can prove to the competent market surveillance authorities, based on recognised performance evaluation methods and 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. Pending the emergence of recognised performance evaluation methods and benchmarks, and as specified in Measure 4.2, Signatories may demonstrate compliance through documented internal testing against the requirements in Commitment 3, subject to review by the competent market surveillance authorities.

The marking techniques may be implemented at different stages of the value chain (e.g., by the provider of the AI system itself or by an upstream model provider) and may also be provided by third parties (in particular, technology providers specialised in marking techniques and detection mechanisms for AI transparency). Signatories may rely on those third-party technical solutions without prejudice to the Signatories' own responsibility under the AI Act and the Code to ensure that the outputs of their AI systems are suitably and compliantly marked.

Sub-measure 1.1.1: Digitally signed metadata

If content is generated, manipulated or exported in a data format that supports attaching metadata (e.g., an audio, image, video, or containerised text), Signatories will record information in the metadata on whether the content is AI-generated or manipulated.

All recorded information will be digitally signed and time-stamped (on systems where time information is available) 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, except in circumstances where the deployment context does not permit the secure handling or provisioning of the key, for example in a local deployment scenario.

Signatories are encouraged to implement richer metadata in accordance with Measure 1.3., , without including privacy-sensitive or business-sensitive information. In cases where such information is strictly necessary or inserted upon request of an end-user, it will be compliant with applicable EU data protection law and is encouraged to be placed in a metadata placeholder separate from the AI transparency metadata.

Sub-measure 1.1.2: Imperceptible watermarking

Signatories will ensure that AI-generated or manipulated content is marked with an imperceptible watermark, with the exception of very short text. For free-form text longer than 200 tokens, watermarking still needs to be applied, even though it may have lower reliability compared to that of watermarking very long text. To compensate for the potential lower reliability of watermarking for free-form text, access to the corresponding detection solution may be restricted to verified expert users as detailed in Sub-measure 2.1.2.

The imperceptible watermark will be embedded within the content in a manner that is difficult for it to be separated from the content. The watermark is intended to serve as a robust mechanism to complement the digitally signed metadata under Sub-measure 1.1.1.

The state of the art provides several strategies to embed watermarks in AI-generated or manipulated content, for instance, applying the watermarking technique once the content has been generated or manipulated (i.e. 'post-hoc watermarking') or having the watermark introduced during the inference operation of the generative AI system (i.e. 'model watermarking'). Signatories who provide generative AI models are encouraged to implement watermarking at the model level and to enable the smooth integration of the watermarking at the AI system's inference process, to facilitate compliance of downstream providers of generative AI systems built on those models, in particular in a manner that helps these downstream providers meet the quality requirements specified in Article 50(2) AI Act and in Commitment 3 and that enables them to demonstrate compliance in line with Commitment 4.

Sub-measure 1.1.3: Fingerprinting or logging (optional)

Where appropriate 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 solutions for AI-generated or manipulated

content which allow for checking whether content has been generated or manipulated by their AI system. For example, direct logging may be appropriate for text content, whereas fingerprinting approaches may be preferable for audio and visual content.

The use of such techniques may be a supplementary solution to meet the quality requirements specified in Article 50(2) AI Act and in Commitment 3 or to help reduce the cost of the detection mechanism outlined in Commitment 2. However, relying on fingerprinting or logging alone is not considered sufficient to meet the quality requirements specified in Article 50(2) AI Act and in Commitment 3.

Where Signatories implement fingerprinting or logging, they will ensure that it is limited to the output data and that it 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 and how data is stored, accessed and how long it is retained, with clear procedures for limited retention periods, access controls, security and other safeguards for data minimisation, as well as secure and immediate deletion of logs containing output data after its intended purpose has been fulfilled.

This Sub-measure is confined to the Signatory's own design and implementation of the AI system and shall in no way be interpreted as creating a general commitment to log, monitor, or retain prompts for the AI-generated or manipulated content or for user interactions.

Measure 1.2: Non-removal of markings

Without prejudice to the robustness requirement in Commitment 3, Signatories will make best efforts to preserve metadata markings on input data and content generated or manipulated by their AI system by applying the following cumulative measures:

a) Signatories will, to the extent technically feasible and recognisable as per open standards, retain, and abstain from intentionally altering or removing, existing metadata markings, where such content is used as input and subsequently transformed by their AI system into an output. This does not affect good faith, legitimate processing where the modification, transformation or replacement of existing metadata is necessary to maintain accurate and functional information following downstream processing or specific legitimate processing, such as for security audits and research purposes.

b) While Signatories recognise that downstream compliance and enforcement cannot be guaranteed, they will nevertheless include in the acceptable use policy, terms and conditions or the documentation accompanying their generative AI system a prohibition of the intentional removal of or tampering with metadata markings by deployers or any other third party. Exceptions to such a prohibition are the legitimate purposes referred to in point a). For AI systems and models provided under free and open-source licenses, it is sufficient for Signatories to alert users to this best practice in the documentation accompanying the AI system or model, without prejudice to the free and open-source nature of the license.

Measures specified in point a) above do not imply the responsibility of the Signatory for third-party metadata markings, nor for compliance of third parties with the conditions specified in point b).

Furthermore, Signatories will neither place or make available on the market, nor promote or advertise the use of tools whose purpose is to circumvent the machine-readable markings added to the AI-generated or manipulated content for transparency.

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

Measure 1.3: Transparency of the provenance information (optional)

Signatories are encouraged to incorporate richer information in the metadata pursuant to Measure 1.1.1, to provide additional context and thereby contribute to increased integrity and trust in the information ecosystem. In particular, Signatories are encouraged to consider relevant standards to provide further information about the origin of AI-generated or manipulated content across workflows, where technically feasible.

Signatories are encouraged to add or record relevant content provenance information within content generated or manipulated by their AI systems, in particular the name of the AI system, the company name of the AI provider, and a timestamp indicating when the content was generated or manipulated. Additionally, Signatories may also include the underlying AI model identifier and version number.

For AI-manipulated content, Signatories are encouraged to record information about the type of operations performed by their AI system to modify the content (e.g., object removal). Multiple discrete operations carried out by the AI system are recommended to be encoded into a single metadata marker to limit complexity and reduce burden.

Measure 1.4: Functionality for perceptible markings (optional)

In order to facilitate compliance of deployers of generative AI systems with their obligation to disclose deep fakes and AI-generated or 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 to implement an integrated option that allows deployers and other users to directly apply at their own discretion – upon generation of the output – a perceptible label in consistency with the Commitments and Measures in Section 2 of the Code.

Signatories are also encouraged to follow harmonised UX standards, to the extent possible, in an interoperable manner with existing standardised content management systems and workflows of online platforms and media publishers.

Signatories are further encouraged to implement other supporting measures for displaying labels and metadata that enable deployers and providers of online platforms, search engines 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 Markings of AI-generated or Manipulated 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 provide the means to enable the detection of the machine-readable markings present in audio, image, video or text content, or a combination thereof, as generated or manipulated by their AI system. Signatories also commit to ensure that the detection result is provided to the natural persons concerned in a clear, distinguishable and accessible manner.

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

Measure 2.1: Detection mechanisms for markings

Signatories will make available a detection solution, composed of one or more detection mechanisms, to enable deployers, users of their generative AI system, third-party integrators, 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 marking technique(s) implemented pursuant to Commitment 1. Signatories will ensure that the detection solution, in conjunction with the associated marking solution, meets the level of effectiveness, interoperability, robustness, and reliability required by Article 50(2) AI Act, Commitment 3 of the Code, and the following Sub-measures.

Sub-measure 2.1.1. Making the detection solution available

The detection solution may be made available in the Union as one or more of the following: (i) a public, ideally standardised, specification allowing any third party to implement a detection mechanism; (ii) a piece of software (e.g., a standalone executable or library); (iii) a cloud-based service accessible to users in the Union through an API. When providing a public specification under point (i), Signatories are encouraged to provide a reference implementation. When providing a piece of software under point (ii), Signatories are encouraged to ensure that the version is portable on most operating systems (including, where technically feasible, mobile phone systems) and that the hardware requirements are limited, if any at all, to facilitate the local execution of the detection mechanism(s) on any general-purpose device.

Signatories will ensure that their detection solution includes a detection mechanism for each marking technique implemented in their generative AI system pursuant to Commitment 1. Similarly to the marking techniques in Measure 1.1, Signatories may rely on shared or third-party detection mechanisms, for instance provided by providers of AI models that apply

markings at the model level or by third-party technology providers specialised in transparency marking techniques and detection mechanisms. Such reliance is without prejudice to the Signatories' own responsibility to ensure compliance with Article 50(2) and (5) AI Act and with this Section of the Code. In order to facilitate compliance by downstream providers of generative AI systems, Signatories who are also providers of generative AI models are encouraged to provide detection mechanisms for the content generated or manipulated by their models prior to the model's placement on the market.

Signatories will make the detection solution available free of charge. Signatories with fewer than 1,000,000 monthly users of their generative AI system, whose detection solution incurs substantial operational costs (in particular when it is provided for detection of watermarks as a cloud-based service), may charge a fee for the use of their detection solution (which will be reasonable, fair and proportionate to the overall operational costs of that solution) in cases where requests from a single user exceed a reasonable threshold due to large volume of requests. All Signatories will always provide free access to their detection solution, without any restriction on the volume of requests, to competent market surveillance authorities and other regulators, law enforcement authorities, media, fact-checkers, trusted flaggers, independent researchers, educational and research institutions, and civil society organisations.

Signatories are encouraged to collaborate with relevant actors within the ecosystem to make their detection solution directly available within distribution and communication platforms.

Sub-measure 2.1.2: Access to the detection solution

Signatories will ensure access to their detection solution through a user interface appropriate for the audience of end-users that may eventually be exposed to the content generated or manipulated by their AI system. Where the general public may be exposed to the AI-generated or manipulated content, the detection solution will accordingly be made available to the general public. However, where there are safeguards to ensure that only a limited number of natural persons will be exposed to the AI-generated or manipulated content (e.g., AI systems in professional settings), and where there are effective safeguards to prevent further dissemination of the content to the general public, access to the detection solution may be restricted to those natural persons.

In their detection solution, Signatories may restrict access to detection mechanisms associated to watermarking techniques for free-form text to the extent that they have a lower level of reliability and robustness and that they may produce misleading or low-confidence results. However, the results of such detection mechanisms may still provide valuable information to verified expert end-users with a legitimate need, including competent market surveillance authorities, other regulators, law enforcement authorities, media, fact-checkers, trusted flaggers, independent researchers, educational and research organisations, and civil society organisations. Therefore, Signatories will ensure that access for these expert end-users is granted subject to appropriate access controls and safeguards. Any restriction to the access will be limited in time until more reliable and robust detection mechanisms have emerged and have been adopted as the state of the art for detection mechanisms for the watermarking of free-form text evolves.

Signatories will ensure that the detection results by their detection solution are provided in accordance with Measure 2.3 below.

Moreover, Signatories will ensure that the detection result of specific content submitted for detection can be downloaded upon request in a digitally signed format, including at least a hash

of the content submitted for detection, a URL or other identifier of the detection solution, and a timestamp.

Sub-measure 2.1.3: Personal data protection, privacy and security

Signatories will ensure that the detection is compliant with EU privacy and data protection law.

Without prejudice to applicable EU law, including EU privacy and data protection law, for any detection solution provided as a software or a service (fully or in part) that requires uploading content, Signatories will ensure:

(a) That the content submitted for detection is processed as privacy sensitive and processed in compliance with EU privacy and data protection law.

(b) The confidentiality and integrity of the content, while in transit and at rest.

(c) Data minimisation in the provision of the detection service, namely that the detection service does not collect, retain, or otherwise process personal data from users of the detection solution, or content submitted to it, beyond what is necessary to deliver the detection functions, except where necessary and based on a valid legal basis for security and abuse prevention, for ensuring the detection solution works effectively, opt-in user feedback or service improvement based on user preferences. Where technically feasible, Signatories are encouraged to apply privacy-preserving technologies and only require the uploading of a derivative of the content that does not give access to the semantics of the content.

(d) That the content submitted for detection is processed for the sole purpose of detecting the markings and is not processed for any other purpose except where necessary for the other purposes specified in point (c) above.

(e) That the content is stored only for the duration of the detection and is permanently deleted immediately thereafter (i.e. with a 'zero retention' policy) provided that minimal data (such as traffic logs) may be retained for a limited period on a valid legal basis solely to ensure security and prevent abuse. In particular, Signatories will not retain a verbatim copy of the content and will delete it immediately after the detection.

(f) That relevant security safeguards, appropriate to the risk, are in place to ensure the ongoing confidentiality, integrity, availability, and resilience of the detection service.

(g) Compliance with all other applicable requirements under EU privacy and data protection law, including the requirements governing transfers of personal data to third countries under EU data protection law.

(h) That best practices in cybersecurity are implemented to protect access to the detection solution in cases where uncontrolled access to the detection solution may compromise the security of the underlying marking techniques and detection mechanisms.

Sub-measure 2.1.4: Retirement of a detection solution

Signatories are entitled to retire a detection solution as long as it is replaced by an alternative detection solution. Signatories will ensure that any such alternative detection solution yields the same or better detection capabilities and is backward compatible, enabling the detection of previously marked content.

Where a Signatory has developed its own detection solution and stops providing that solution without replacement for any reason (e.g., business closure), it will make that solution available

to the competent market surveillance authorities, subject to confidentiality and security safeguards to enable them to detect, where necessary, legacy content generated or manipulated by their AI system.

Measure 2.2: Forensic detection mechanism (optional)

To complement the marking techniques specified in Commitment 1 and contingent upon their capacities, Signatories may include as part of their detection solution, as an additional optional measure, forensic detection mechanism to detect content generated or manipulated by their AI system or underlying model for which marking has been stripped. Where Signatories implement such a forensic detection mechanism, they will ensure that the privacy and security requirements as specified in Sub-measure 2.1.3 are satisfied and that the disclosure of the detection result is clear and accessible as specified in Measure 2.3, and make best efforts to ensure that their forensic detection mechanism aligns with the state of the art as specified in Commitment 3.

At the time of publication of the Code, forensic detection mechanisms were not deemed mature enough to comply with the quality requirements set in Article 50(2) AI Act and in Commitment 3. As long as no better forensic detection mechanism has emerged under the state of the art, Signatories who implement forensic detection mechanisms as an optional measure may restrict access to the detection results provided by such forensic detection mechanisms to verified expert users, in line with the exception for detection mechanisms for free-form text watermarking described in Sub-measure 2.1.2.

Measure 2.3: Clear and accessible disclosure of detection results

In order to fulfil their obligation under Article 50(5) AI Act, Signatories will ensure that the detection results provided by their detection solution are presented in a way that is clear and easily comprehensible to natural persons exposed to the content generated or manipulated by their AI system and who want to verify its origin.

Signatories will ensure that detection results indicate whether they are based on a metadata marking, a watermark marking, forensic detection or other techniques, to the extent technically feasible.

Where additional information is available in the watermark or in the metadata, Signatories will incorporate such information into the detection results. Signatories may also include further information within detection results, such as the proportion of the content that has been generated or manipulated by their AI system and/or the localisation of the AI changes in the content. For detailed detection results, Signatories are encouraged to consider tiered access to avoid overwhelming end-users with possibly confusing information, especially when the end-user of the detection solution is a member of the general public lacking specialised knowledge.

Where applicable, Signatories will ensure that human interfaces used to present detection results are accessible to persons with disabilities in compliance with applicable accessibility requirements under Union law, in particular those laid down in Directive (EU) 2019/882 (the European Accessibility Act) and Directive (EU) 2016/2102 (the Web Accessibility Directive). 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 and detection solutions (optional)

Signatories are encouraged to ensure that documentation and other relevant information (excluding business confidential information and trade secrets) is provided to deployers and other expert users to support them in making informed decisions on what marking and associated detection solutions they may use, including helping them to understand how to access detection solutions, how to perform detections, and how to interpret 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 laypersons and end-user needs, including where the AI systems serve populations with lower AI literacy or in sensitive contexts (e.g. educational contexts or contexts involving young or elderly users). These resources may be developed by the Signatories themselves or created jointly through efforts involving other providers or by organisations or initiatives they belong to.

This measure will 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 exposed to the content, and the size and resources of the AI system 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 transparency, and to support EU-level initiatives to foster a consistent understanding of provenance and detection across Member States.

Commitment 3: Measures to Meet the Requirements for Marking and Detection Solutions

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 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 that their marking and detection solutions comply with the quality requirements described in the following measures, holistically across all marking techniques employed rather than for each technique individually. Signatories commit to assess and demonstrate compliance with the requirements in accordance with the testing, verification, and compliance processes specified in Commitment 4 prior to placing their generative AI system on the market or putting it into service, and throughout its lifecycle.

Signatories commit to strive to achieve the best possible balance between effectiveness, interoperability, robustness and reliability of the marking and detection solutions as described in Measures 3.1 to 3.4 below. Operational constraints may impose limitations on computational

time and resources, costs of implementation, as well as on scalability to very large or very small content.

Measure 3.1: Effectiveness

Signatories will implement technical marking and detection solutions which, in conjunction, are fit-for-purpose and capable of enabling natural persons to distinguish content generated or manipulated by their AI system, thus contributing to the trust and integrity of the information ecosystem.

Those solutions will be considered effective when natural persons can access and understand the meaning of the detection results. There is no quantitative evaluation metric linked to this Measure; instead, it requires a user-based assessment of the detection functionality. Signatories are encouraged to publish specifications to guide downstream stakeholders on how to display the detection information output by their detection solution. Signatories may conduct relevant panel studies, to the extent relevant to their application use case, to check how well natural persons exposed to the detection results understand them.

Measure 3.2: Reliability

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

Reliability refers to the capability of the marking and detection solutions to accurately identify and distinguish the origin of AI-generated or manipulated content from other content. The reliability assessment consists of two components: (i) how accurate the detection solution is with regard to detecting the marking in nominal conditions where there is no alteration to the content, and (ii) how the accuracy of the marking and detection solutions varies with respect to the length, size, diversity, and semantics of the content.

Signatories will measure the accuracy of the detection of AI-generated or manipulated content by using relevant metrics for the type of markings, such as the error rate of the detection. Low error rates will be demonstrated on a variety of samples, including both content generated or manipulated by their AI system with their own marking and other content (possibly including content generated or manipulated by other AI systems). The data for the assessment of reliability shall, to the extent possible, not have been used during the training and development of the AI system.

Signatories will ensure these accuracy measurements are performed and reported with content intended to be generated or manipulated by the AI system of various length, size, diversity, and semantics, in order to demonstrate that the performance of the implemented marking and detection solutions generalises across diverse contexts and use cases.

Measure 3.3: Robustness

Signatories will implement marking and detection solutions which, together, maintain intended performance levels under varying conditions, covering both common alterations and adversarial attacks, to the extent technically feasible, and in alignment with the state of the art. This requirement does not apply to AI systems that are exceptionally subject only to one layer of metadata marking as specified in Measure 1.1.

First, Signatories will ensure their marking and detection solutions are robust to typical processing operations that may introduce varying degrees of distortion based on underlying parameter settings. Typical processing operations include:

a) in-place modifications, e.g. noise addition / removal, filtering, (re)compression, screenshot / screencasting, voice enhancement, lexical substitution, homoglyphs, change of file format, etc;

b) desynchronisation mechanisms, e.g. mirroring, cropping, up/downscaling, rotation, aspect ratio change, pitch shifting, time stretching, characters insertion / deletion, paraphrasing, translation cycles, etc;

c) survival of the analogue hole, e.g. print-and-scan (and optical character recognition), audio playback and recording, screen camcording, etc.

Moreover, Signatories will ensure that detection is robust to localised changes to the content, for example if users decide to blur faces for privacy reasons prior to submission of the content to the detection solution of the Signatory.

Second, Signatories will ensure that adversarial robustness of their marking and detection solutions is assessed in terms of resilience to malicious behaviour such as copying, removal, regeneration, and modification attacks on the markings. Signatories will measure robustness to malicious behaviour by the ability of the marking and detection solutions to verify marking integrity, i.e. whether a mark has been tampered with or modified to misrepresent the origin of the content. Signatories will choose the considered malicious behaviour as plausible real-world threats based on the type of content, the type of mark and the context of deployment and dissemination of the content. Signatories are encouraged to apply standard cybersecurity practices, such as rate limits, to prevent and counteract malicious use and attacks against their marking and detection solutions. Signatories are encouraged to frequently update their threat assessment to keep up to date with changes in the threat landscape.

To assess the robustness to typical processing operations and malicious behaviour, Signatories will use the same performance metrics as for the reliability requirement in Measure 3.2.

Measure 3.4: Interoperability

Signatories will implement marking and detection solutions that operate seamlessly across multiple systems, actors, contexts and technical implementations to enable detection of AI-generated or manipulated content, regardless of the marking technique deployed in different AI systems, to the extent technically feasible.

At the time of publication of this Code, relevant interoperability standards and/or best practices are yet to be developed, except for digitally signed metadata. Therefore, a staged implementation of interoperability requirements will be adopted as follows.

In the initial stage as of the entry into application of Article 50(2) AI Act, Signatories will:

(a) adopt relevant established standards or best practices to implement metadata marking and detection, thereby enabling open detection of such markings;

(b) provide public information to the Commission and relevant stakeholders on how to integrate and access detection solutions for the other type(s) of marking technique(s) that they have implemented in their AI system. To facilitate dissemination of such information, Signatories are encouraged to provide relevant information on their websites and to populate publicly available registries that may be created to streamline integrations.

(c) work towards a minimum level of interoperability among detection solutions for watermarking that reduce burden and costs for the Signatories, while ensuring a more seamless access for end-users. More specifically, Signatories will implement an interoperability solution for their detection mechanisms by 2 February 2027 by implementing one or more of the following:

i) a publicly available interoperable industry standard access method that enables detection queries to be routed to relevant detection mechanisms, and detection results to be interpreted by shared or provider-agnostic detection solutions interacting with endpoints implementing such API, including detection solutions that support public signposts but do not necessarily require them. Signatories who implement this approach shall not create unnecessary traffic to other services where such traffic can be avoided by directly reading out a known interoperability solution (like a signpost);

ii) a publicly readable signpost or other interoperable mechanism in the AI-generated or manipulated content that will signal to the public which detection solution to use without having to run all the detection solutions of providers of AI systems. Signatories will take reasonable steps to ensure that the detection mechanism(s) associated to their signpost marking technique(s) or other interoperability solution are made available to the Commission and relevant stakeholders so that they can be integrated in their detection workflows;

iii) a shared provider detection solution by a consortium of Signatories that is accessible and inclusive for any other Signatory to join, including SMEs and SMCs, subject to appropriate safeguards for the protection of confidential technical information. Any such aggregated detection solution shall be agnostic to the Signatories participating in the consortium, capable of detecting outputs generated or manipulated by their generative AI systems, and able to work in an interoperable manner with other detection solutions;

iv) another interoperability solution that achieves interoperability between Signatories comparable to the solutions above.

The interoperability solution will be implemented in a manner that is not detrimental to the robustness of other embedded watermarks and that includes safeguards against misuse risks creation.

(d) To advance towards full interoperability, contingent upon their resources and capacities, Signatories will cooperate in good faith with the Commission and relevant stakeholders towards the development of improved and state of the art interoperability solutions and interoperability standards. In particular, Signatories are encouraged to participate in the task force described in Measure 3.5. Signatories are also encouraged to join and/or support international and European standardisation bodies, as well as consortia initiatives focused on the development of open content marking and detection standards that help to operationalise the measures envisaged in this Section of the Code.

In these subsequent stages, Signatories commit to implement those standards in a timely manner as they emerge, to the extent technically feasible and needed for the application use case of their AI systems, in order to continue to be compliant with the requirement for interoperability under Article 50(2) AI Act.

Measure 3.5: Advancing the state of the art (optional)

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 techniques and detection mechanisms for AI-generated and manipulated content.

Such advances of the state of the art comprise, inter alia:

a) advancing existing marking techniques and detection mechanisms with regard to the achievable level of effectiveness, reliability, robustness, and interoperability;

b) advancing existing interoperability solutions towards higher degree of interoperability;

c) contributing to and promoting the standardisation of the interfaces of marking and detection solutions, and the techniques themselves when possible;

d) developing new marking or detection solutions that are agnostic to the AI provider, for instance detection mechanism(s) that are operational with regard to the quality requirements above across multiple AI systems, including publicly available detection solutions that can be run locally, without reliance on third-party detection services;

e) establishing industry best practices and/or public benchmarks for evaluating marking and detection solutions and the underlying techniques, including contributing to the creation of reference datasets;

f) organising and/or participating in red-teaming exercises to assess the limits of state-of-the-art marking and detection solutions, and the underlying techniques, and to identify new deficiencies.

Finally, Signatories are encouraged to actively participate in the taskforce that will be set up under the Code to facilitate the activities mentioned above, the sharing of good practices, consistency in the implementation of the Measures under the Code, and collaboration between Signatories, other actors in the value chain and relevant stakeholders through regular meetings.

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 compliance, testing, verification and monitoring processes, as specified in the following measures.

Measure 4.1: Compliance process

Signatories will document, implement, and keep up to date, in line with the state of the art, a compliance process that describes at a high-level how they have implemented the different Measures in this Section of the Code to ensure compliance with Article 50(2) and (5) AI Act. This measure will be implemented in a proportionate manner, taking into account the size and resources of the Signatory, in particular with regard to providers that are SMEs and SMCs to minimise burden and simplify compliance.

Signatories may demonstrate compliance through existing processes and compliance documentation 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 may leverage documentation provided by those third parties, without prejudice to the ultimate responsibility of the Signatory to ensure compliance with Article 50(2) and (5) AI Act. Signatories who rely on open-source marking and detection solutions and standards will document the reliance on those standards and tools, as well as additional information, only with regard to the measures in this Section of the Code for which public information is lacking.

Measure 4.2: Testing, verification, and monitoring

Prior to placing their generative AI system on the market or putting it into service, and regularly thereafter, Signatories will test the compliance of their marking and detection solutions with the requirements and the measures specified in this Section of the Code. This will be carried out under conditions that are representative of real-world use, with the objective of ensuring alignment with the evolving state of the art.

Signatories who are downstream providers of generative AI systems may rely on the results of testing performed by an upstream model provider or a third-party provider of marking and detection solutions, so long as those solutions comply with the requirements specified in this measure and without prejudice to the ultimate responsibility of the Signatory to ensure compliance with Article 50(2) and (5) AI Act.

Until recognised performance evaluation methods and benchmarks for marking techniques and detection mechanisms emerge, in particular methods and benchmarks endorsed by the AI Office after consultation of the AI Board and the taskforce described in Measure 3.5, Signatories will test and report on the performances of their solution according to internal benchmarks and industry best practices. Signatories may involve independent experts in the testing of their solutions, especially to evaluate robustness to adversarial attacks through red teaming and/or conduct such testing and evaluation under regulatory supervision in the context of AI regulatory sandboxes, as provided for in Article 57 AI Act.

Signatories will monitor any substantiated compliance shortcomings identified internally or reported by relevant third parties and implement appropriate follow-up corrective actions.

Measure 4.3: Training

Signatories will make proportionate efforts to provide appropriate training to their personnel who have roles relevant to ensuring compliance with Article 50(2) and (5) and Article 4 AI Act. This comprises personnel or external contractors who are involved in the design and development of the Signatories' AI systems, and who are responsible for ensuring that the Commitments and Measures specified in this Section of the Code are implemented effectively. This measure will be implemented in a proportionate manner, taking into account the size and resources of the Signatory, in particular with regard to providers that are SMEs and SMCs.

Measure 4.4: Cooperation with market surveillance authorities

In line with Article 74 AI Act and Article 7 of Regulation (EU) 2019/1020 (the Market Surveillance Regulation), 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. At their reasoned request, where necessary for the exercise of the relevant tasks of the market surveillance authorities, Signatories will provide the documentation specified in Measure 4.1 and access to their marking and detection solutions. 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 transparency across the value chain and that have been assessed by the AI Office and the AI Board to be compliant with the Code, for instance via an EU 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
Analog hole	Refers to a category of content processing operations where the digital content is rendered in an analogue form, typically to be presented to natural persons, before being re-digitised. A simple example is recording digital video content displayed on a screen with a digital camera.
API	Application Programming Interface: a machine-usable interface to an AI system or another software service.
Containerised text	Text that is embedded within a structured file format or a container, where the content is organised according to specific rules defined by the format. The container may include metadata, layout information, and encoding details that influence how the text is stored, displayed, and processed. Examples include text inside PDFs, Word documents, or HTML files.
Desynchronization mechanism	Refers to a category of content processing operations where the position of the digital samples is modified (and also possibly their value). A simple example is applying a mirror effect to a digital image. This definition has to be considered side-by-side with the one for 'in-place modification'.
Detection mechanism for a marking technique	Mechanism for the detection of watermarks or the verification of digitally signed metadata markings that have been purposefully added by a provider of an AI system or a third party (e.g. model provider). Different providers of an AI system may share the same detection mechanism, typically when they use a standardized marking technique or develop a shared detection solution jointly. However, in general, a detection mechanism is specific to a given marking technique, i.e.

	different watermarking techniques will have different detection mechanisms.
Digital signature	A cryptographic signature that enables secure verification of authenticity of the provider and integrity of the signed content.
End-user	A natural person exposed to the content generated or manipulated by the AI system.
Error rate	Relates to the probability of the detection mechanism to misclassify some content as "AI-generated or manipulated". Error rates can be further broken down for improved interpretability, e.g. the false positive rate can negatively affect the trust that can be put in a marking technique and associated detection mechanism.
Fingerprinting	Also referred to as 'perceptual hashing' or 'robust hash'. Content indexing technique that reduces content to condensed descriptors that can be efficiently looked up even if content has been subject to alterations. Fingerprinting has been used in the past for content-based recognition or identification.
Forensic detection	Detection of AI-generated or manipulated content which does not require prior AI marking. For example, a forensic detection mechanism may aim to capture some intrinsic signal signature present in AI-generated or manipulated content that makes it different from other content.
Free-form text	A raw sequence of characters with no enclosing structure, schema, or container format. It is not bound to any metadata, layout rules, or encoding beyond basic character representation. Examples could be the text that is shown on a website or within a chat.
Homoglyphs	Characters that appear (nearly) identical in shape but that originate from different alphabets, fonts, or encodings. As such, they may be exploited to confuse text marking techniques and associated detection mechanisms.
In-place modification	Refers to a category of content processing operations where the value of the digital samples, e.g. image pixels, are modified, but not their locations. A simple example is increasing the contrast of a digital image. This definition has to be considered side-by-side with the one for 'desynchronization mechanism'.
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.
Marking	Addition or embedding of a marker to AI-generated or manipulated content such as a digitally signed metadata or imperceptible watermarking. The purpose of this marker is to facilitate the detection of AI-generated or manipulated content.

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.
Screencasting	Process of digitally recording the visual output displayed on a screen (computer, tablet, smartphone), often with some audio narration, and saving it to a file or streaming it live.
Signpost solution	Openly disclosed marking technique within the content and associated detection mechanism used to indicate which (watermark) detection mechanism shall be employed. This is an interoperability mechanism to streamline the detection workflow and avoid running each content subject to verification against all detection mechanisms employed by providers of generative AI systems whenever metadata is not available. It may be implemented, for instance, as an imperceptible machine-readable mark.
User	Either a deployer within the meaning of Article 3 (4) AI Act or another person that is using the generative AI system of a provider.
UX	"User Experience", i.e., the way a user interacts with and perceives user-sided aspects of a software product.
Very short text	Text that is so short that in many cases it cannot be watermarked, even with a basic level of reliability. At the time of publication of the code, state-of-the-art techniques enable watermarking with at least a basic level of reliability of text as short as 200 tokens, with the expectation that this threshold will decrease as new methods become available. Until then, "very short text" should be understood as text shorter than 200 tokens.
Watermark	Technique that embeds a marker within the content in a manner that is imperceptible (to natural persons) and robust (the watermark is recoverable after content alteration). The watermark is inherently tied to the digital representation of the content (e.g. image pixels, audio samples) i.e. it is not auxiliary metadata. Watermarking has been used in the past for anti-piracy, audience measurement, and proof of ownership.

Section 2:

Labelling deep fakes 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: Labelling deep fakes and 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.

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 that deployers of AI systems that generate or manipulate image, audio or video content constituting a deep fake, or text published with the purpose of informing the public on matters of public interest comply with their obligations under Article 50(4) and (5) AI Act, and

c) to enable the competent market surveillance authorities to assess compliance by deployers who rely on the Code to demonstrate compliance with those obligations under Article 50(4) and (5) AI Act in a consistent, predictable, and uniform manner across the Union.

Recitals

Whereas:

a) **Clear and distinguishable disclosure:** Signatories recognise that advances in generative AI enable deployers of AI systems to generate or manipulate content that looks increasingly realistic, making it more difficult for people to distinguish it from authentic content. This highlights the need for transparency to protect public trust, safeguard democratic discourse, and inform individual decision-making. With the assistance of AI systems, deployers can generate or manipulate deep fakes that resemble existing persons, objects, places, entities or events, falsely appearing authentic or truthful to natural persons, as well as text on matters of public interest that is published without human review and editorial control. Generation and dissemination of such content pose specific risks of deception and manipulation, undermining trust in the information ecosystem and threatening its integrity. Signatories acknowledge that clear and distinguishable disclosure of the artificial origin or modification of such content required under Article 50(4) and (5) AI Act is a necessary safeguard to make individuals aware and inform their decision-making. This safeguard is not aimed at providing individuals with information about the trustworthiness of the content,

but rather at informing them about its artificial origin or the role of the AI system in manipulating the content.

b) **Consistent and effective implementation:** To make such disclosure clear and distinguishable for all natural persons, Signatories recognise the need for a certain level of uniformity in the disclosure methods as laid out in the Commitments of this Section of the Code. Such methods must be effective and aligned with the technical state of the art (including for user experience) to the extent feasible and appropriate for the context of dissemination. In particular, an optional EU icon provides an easy, practical and uniform pathway under the Code for compliance with Article 50(4) and (5) AI Act. At the same time, Signatories recognise the broad transversal scope of Article 50(4) AI Act as covering a multitude of use cases and contexts across sectors, media, platforms and modalities. Signatories acknowledge the need for a minimum level of uniformity and consistency regarding the implementation of the design and placement specifications for the icon or equivalent labels. At the same time, Signatories may further detail how to implement the measures envisaged in this Section of the Code in sector-specific good practices compliant with Article 50 AI Act and this Section of the Code.

c) **Accessibility:** Signatories recognise the importance of ensuring accessible disclosure of AI-generated or manipulated deep fakes and text published on matters of public interest within the scope of Article 50(4) AI Act, by taking into account the needs of different groups of vulnerable people. Disclosure should be designed to ensure it is easily perceivable and understandable by all natural persons, including persons with disabilities, children and the elderly.

d) **Awareness and accountability:** Signatories recognise that internal processes, literacy and awareness measures, and review mechanisms are essential for ensuring effective disclosure of AI-generated or manipulated deep fakes and text published on matters of public interest within the scope of Article 50(4) AI Act. To achieve this, Signatories commit to label such AI-generated and manipulated content and to disclose it at the latest at the first exposure based on the design and placement specifications in this Code. Considering Signatories' available organisational resources and technical capacities, this also includes proportionate internal processes and measures which aim to ensure that individuals are informed about the artificial nature of the content to which they are exposed, including by making publicly available information about the use by the Signatory of the EU icon or equivalent labels and their meaning.

e) **Promoting an ecosystem of transparency and innovation:** Signatories acknowledge that compliance with Article 50(4) AI Act is not a check-box exercise but aims to create an ecosystem of transparency that ties into organisational practices, technological developments, and individuals' awareness. It has the shared goal of empowering individuals to make informed decisions by enabling them to understand the origin and nature of synthetic content. In such an ecosystem, Signatories recognise that the Code allows for continuous developments in a fast-changing technological landscape and in light of the transversal scope of Article 50(4) AI Act. Key factors in such a structure of the Code are to build an ecosystem that preserves uniformity and recognisability, while allowing for new evidence to be gathered and adapted into good practices that support the Code and compliance with Article 50(4) and (5) AI Act.

f) **Additional safeguards under other Union and national law:** Signatories acknowledge that the transparency obligations under the AI Act apply in parallel to, and do not affect, other legal obligations or requirements that may apply to the creation, distribution or use of AI-generated or manipulated deep fakes and text published on matters of public interest under applicable Union legislation (e.g., on data protection, consumer protection, digital services, intellectual property, media law, 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 image, audio or video content constituting a deep fake[1] or text published 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", "deep fake" or "published text" implies content that is i) AI-generated or manipulated and that qualifies as a deep fake under Article 3(60) AI Act (referred to as "deep fake"), or ii) AI-generated or manipulated 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").

Considering the diversity of deployers and contexts in which deep fakes and published text may be created and disseminated, Signatories and the associations of which they are part may jointly establish common good practices for their sector or industry, which could provide further detail regarding implementation of the Commitments and Measures specified in this Section of the Code. Any such good practices should be fully aligned with the measures set out below, without preventing Signatories from tailoring the disclosure practices to their specific sectors to account for sector specificities and existing rules, nor from providing additional transparency, where relevant. To ensure transparency, accountability and to enable monitoring by competent market surveillance authorities, sectoral good practices established in accordance with this paragraph will be made publicly available online.

In this Section, the terms "will", "encouraged" and "may" should be interpreted in the following way:

1. **'will'** is used for mandatory measures under the Code that need to be met to be compliant with Article 50(4) or (5) AI Act and that will be monitored by competent market surveillance authorities;
2. **'encouraged'** is used for optional measures under the Code that are not legally required and are purely voluntary but are nevertheless recommended;
3. **'may'** is used for optional measures under the Code that are not legally required and are purely voluntary or provide deployers with flexibility on how to implement the respective measure or commitment.

[1] A deep fake is defined in Article 3(60) AI Act as AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful.

Commitment 1: Disclosure of 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.

To fulfil their obligations under Article 50(4) and (5) AI Act, Signatories commit to ensure consistent and effective disclosure of the artificial origin of deep fakes or published text.

Signatories commit to implement such disclosure through the available EU icon provided in Annex 1 of the Code or through an equivalent icon or label that complies with the design and placement specifications as described in the following measures.

Signatories may integrate the icon or equivalent label into their existing disclosure practices under applicable sectoral Union and national legislation or professional codes of conduct, to the extent that those practices comply with this Section of the Code and Article 50(4) and (5) AI Act.

Signatories recognise that the act 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 freedoms, including applicable legal requirements regarding obtaining consent of the depicted persons or rightsholders.

Measure 1.1: Design specifications

Where visual disclosure is possible, Signatories will implement the following design specifications:

a) The icon or equivalent label will comprise, as the main visual element, the capitalised acronym "AI" in the English language (e.g., an 'AI' icon), unless use of English is incompatible with applicable national laws on the use of languages in commercial or administrative matters in which case the acronym may be disclosed in the national language. The letters in the acronym will have the same vertical dimension. If resized, the proportions of the letters must be preserved.

b) Additionally, Signatories are encouraged to supplement the acronym in the icon with information regarding whether the deep fake or published text is manipulated or generated with AI in an (interactive) second layer where this is technically implementable, in the icon or next to the icon (e.g., text indicating “modified” or “generated”, as illustrated in Annex 1). Signatories are furthermore encouraged to disclose in the (interactive) second layer what has been modified by the AI system (e.g., text or pictogram describing that a face has been altered).

c) The icon or equivalent label may appear in different sizes depending on the context, as long as Signatories ensure that the disclosure is clear and distinguishable to natural persons. The icon or equivalent label may be expressed in different styles (e.g., contrast ratio, colour, or typography), as long as it remains clear, accessible, and distinguishable, i.e., readable and recognisable to natural persons.

An EU icon following the design specifications for visual disclosure is publicly available for everyone to use freely (see Annex 1).

Where visual disclosure is not possible (e.g., audio-only content), Signatories will implement the following design specifications:

a) The disclosure will include, at the beginning of the deep fake itself, a short audible disclaimer in plain and simple natural language, either in the same language as the content or in English, disclosing the artificial origin of the audio deep fake in a perceivable manner.

b) Where appropriate, the audio disclaimer will be complemented with information regarding whether it is AI-generated or manipulated. Signatories are furthermore encouraged to disclose what has been modified by the AI system.

Alternatively, Signatories may develop and rely on other forms of audible disclosures (e.g., an earcon) ahead of the development of a common EU-wide audio solution by the Task Force, described in Measure 1.3, if the deployment of such an audible disclosure solution is accompanied by appropriate awareness raising measures (e.g., public information campaigns, repeated explanatory notices or disclaimers) to ensure broad understanding of the solution by the natural persons exposed to the deep fake audio.

When applying the design specifications 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 vulnerable user categories such as children and the elderly), and the potentially sensitive nature of the context in which the content is used and disseminated (e.g., in the financial, medical, education or other sensitive sectors).

Signatories will ensure accessible disclosure with respect to different content modalities and contexts in accordance with applicable Union law, in particular Directive (EU) 2019/882 (the European Accessibility Act) and Directive (EU) 2016/2102 (the Web Accessibility Directive). This includes but is not limited to application of:

a) audio descriptions or alternative cues for visual disclosure elements;

b) tactile and haptic cues for audio-only content, considering the needs of end-users with hearing impediments (e.g., a vibration alert before audio playing identifying that the audio contains deep fake content);

c) high contrast icons or equivalent labels and screen-reader compatibility, including for end-users with colour vision deficiencies;

d) detectability of the icon or equivalent 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 the applicable accessibility requirements and such standards or guidelines.

Signatories that are also providers of online platforms or online search engines, including designated very large online platforms or very large online search engines subject to the obligation to put in place risk mitigation measures under Article 35(1) Digital Services Act (DSA), are encouraged to use the EU icon or another equivalently compliant solution to enable deployers of AI systems to implement their labelling obligation in a manner compliant with this Section of the Code and Article 50(4) AI Act (e.g., by making disclosure solutions available in the platform upload interfaces for deployers).

Measure 1.2: Placement specifications

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 and considering the dissemination of published text and deep fakes in particular across online platforms or search engines in both offline and online environments, Signatories will apply the placement specifications described in this measure.

Sub-measure 1.2.1: Overarching principles for placement specifications

Signatories will apply the following overarching principles applicable to all content modalities and contexts, with the exception of deep fakes that are part of artistic, creative, satirical, fictional or analogous works that are subject to specific disclosure regime specified in Commitment 3:

a) Considering the content format and dissemination context, the icon or equivalent label will be placed in an appropriate and perceivable manner that ensures immediate recognition by natural persons without requiring user (inter)action or sustained attention.

b) The icon or equivalent label will remain visible at least for a sufficient duration to be noticed under normal exposure conditions to ensure perceivability.

c) The icon or equivalent label will be directly embedded into the content, unless equivalent alternatives to an embedded icon are available (e.g., a user interface overlay that for natural persons appears to be on the content), as long as Signatories follow the placement specifications in this measure and, to their best effort aim to ensure disclosure of the deep fake or published text takes into account the distribution and dissemination chain of the content.

d) The icon or equivalent label will be clearly perceivable and distinguishable at the latest at the time of first exposure of a natural person to the deep fake or published text (e.g., by maintaining sufficient spacing to other overlay elements, disclaimers, and sound/visual on-screen/on-display elements and remain visible against any background).

Signatories are encouraged to collaborate on a best effort basis with actors whose services or products they use to further distribute or disseminate the deep fake or published text (e.g., publishers, online platforms or retail) to preserve the icon or equivalent label with the objective

of ensuring that the required disclosure accompanies the deep fake or published text throughout its further distribution, whether online or offline and across the entire content creation, distribution and dissemination chain. This encouragement also applies to disclosures pursuant to Commitments 3 and 4 of this Section of the Code.

Sub-measure 1.2.2: Placement specifications where visual disclosure is possible

In addition to the overarching principles, Signatories will implement the following placement specifications where visual disclosure is possible:

a) The icon or equivalent label will appear in an appropriate place where no intervening overlay elements exist (e.g., in the top right corner of an image or video deep fake).

b) To account for situations where different natural persons may be exposed to deep fake video at different moments (e.g., live content) and to account for downstream use (e.g., screenshots and clipped fragments), Signatories will display the icon or equivalent label at the beginning of the video as well as, where possible, at regular intervals throughout the video and, at a minimum, after interruptions (e.g., after commercial or advertising breaks). Where possible and appropriate depending on the context, Signatories are furthermore encouraged to display the icon or equivalent label throughout the deep fake video or during the deep fake part of the video, where the deep fake segment does not necessarily span the entire duration of the video.

c) Where a deep fake is exclusively used as part of a closed internal professional context (e.g., for the purpose of training or informing employees), the disclosure may be placed in the user interface, physical setting or any other appropriate medium readily available to the natural persons exposed to the deep fake, informing them before being exposed to the deep fake, as long as it is clear and distinguishable for the natural persons involved in the professional setting.

d) For visual deep fakes, audible disclosures may only be implemented as an additional disclosure method and will always be accompanied by visual disclosures (e.g., where audio is part of the user experience or as appropriate for accessibility purposes).

e) For audio deep fakes, when a screen is available, an additional visual disclosure based on the icon or equivalent label will be made available (e.g., when a screen is available in a car or on a smartphone display), in addition to the audible disclaimer.

f) For published text, Signatories will place the icon or equivalent label, 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, consistent, and distinguishable for the end-user. Where appropriate, Signatories may label only that part of the text which is AI-generated or manipulated. For short text (single words or brief phrases), where labelling the text outputs would degrade readability and usability, Signatories must still carry out the labelling but may ensure disclosure through a contextual notice in the user interface (e.g., an indicator adjacent to the output, or disclosing that AI is used in the beginning of a user exposure session or when first time interacting with the content).

Sub-measure 1.2.3: Placement specifications where visual disclosure is not possible

In addition to the overarching principles, where visual disclosure is not possible, Signatories will include an audible disclaimer at the latest at the time of the first exposure to the deep fake. Additionally, audible cues may be used (e.g., tones or earcons) to support recognition.

In order to cater for downstream use (e.g., clipped audio fragments) and for long-form or live exposure, Signatories will, where appropriate, supplement the initial audible disclaimer at the beginning of the content with reminders at regular intervals (e.g., disclaimers, tones or earcons) for the entire duration of the audio deep fake to ensure ongoing awareness, and at minimum after interruptions (e.g., by advertisements).

Measure 1.3: Voluntary participation in a task force under the Code

Signatories are encouraged to support the work and activities of a dedicated task force under the Code aimed at advancing the further development and usability of the EU icon as a minimum state-of-the-art implementation. The task force will, inter alia, aim to:

a) further promote the uptake and development of the EU icon (incl. non-visual/audio-only versions of the icon);

b) update the EU icon to ensure state-of-art disclosure at all times including in line with user experience (UX) standards;

c) assess feasibility and work towards advancing technical solutions, including the development of an interactive second layer linked to the EU icon or other equivalent labels, which provide opportunities for the integration of the EU icon or equivalent label into online content dissemination services to provide end-users with additional information (e.g., type of AI-involvement in a deep fake or published text). This interactive second layer may be implemented to the extent feasible and practical, and in alignment with the technical marking solutions described in Section 1 of the Code, to ensure robustness of the information provided in the icon and the accompanying information;

d) work towards identifying common methods for disclosing the type of AI involvement in the case of AI manipulated deep fakes or published text, with a view to advancing consistency in the implementation among Signatories;

e) provide a forum for Signatories to exchange good practices across sectors and possibly to develop sectoral good practices regarding the disclosure of deep fakes and published text in line with this Section of the Code.

Signatories and relevant stakeholders from various sectors, including but not limited to independent researchers, civil society organisations and national competent authorities, are encouraged to participate in the task force.

Commitment 2: Internal Processes

To effectively fulfil and demonstrate compliance with the obligations under Article 50(4) and (5) AI Act and the Commitments and Measures specified in this Section of the Code, Signatories commit to put in place or maintain internal processes, awareness measures and review mechanisms as specified in the measures below and proportionate to their size and available resources.

Measure 2.1: Internal compliance process

Proportionate to their size, available organisational resources, and capacities, Signatories will put in place or maintain appropriate internal compliance processes and documentation that specifies how they implement the disclosure obligations using the icon or equivalent label. Such documentation may include a general description and representative, concrete and real

examples of how disclosures are implemented in practice in accordance with Commitments 1, 3 and 4.

Signatories are encouraged to publish information on the disclosure solution they apply in a clear and accessible manner. If Signatories use the publicly available EU icon under Commitment 1, they may do so by using the information provided with the EU icon. If a Signatory chooses to apply their own icon or equivalent label, a similar description of the icon or label is encouraged to be made publicly available. This may include images of the icon or equivalent label(s) in use, as well as explanations of the meaning of each label. Where an icon or equivalent label carries a different meaning depending on the context, it is encouraged that any variations are explained.

Where appropriate and particularly in cases of regular use of AI systems to create deep fakes or published text, Signatories will put in place or maintain a process to ensure and verify that the design and placement specifications are implemented properly to mitigate risks of non-labelled or incorrectly labelled content.

Media service providers within the meaning of Article 2(2) of Regulation (EU) 2024/1083 may comply with this Measure by applying their existing procedures and established professional standards, to the extent that those practices comply with the measures within this Commitment and Article 50(4) and (5) AI Act.

Measure 2.2: Awareness and literacy

Signatories will make efforts proportionate to their size, available organisational resources, and capacities to ensure awareness of the disclosure obligations under Article 50(4) and (5) AI Act among their personnel, including employees, and external contractors directly involved in the implementation of disclosure 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, cases involving artistic, creative, satirical, fictional or analogous work, accessibility considerations, and procedures for correcting missing or incorrect labels where these have been identified.

AI literacy related efforts should be proportionate to the size and available resources of the Signatory, and applied as appropriate to the roles, technical knowledge, experience, and education of personnel involved in creating, modifying, and disseminating relevant content subject to the labelling obligations. Signatories remain free to determine the training formats and their frequency.

Measure 2.3: Review, feedback and cooperation with authorities

Signatories will support the effective implementation of the design and placement specifications through internal review and external feedback.

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

Signatories will review cases that have been reported in a substantiated manner as mislabelled or incorrectly labelled and take measures to remedy cases of non-compliance with Article 50(4)

and (5) without undue delay. Signatories will cooperate with competent authorities in accordance with applicable Union and national laws (e.g., national market surveillance authorities).

Commitment 3: Disclosure for Artistic, Creative and Similar Works

In accordance with Article 50(4) AI Act, Signatories commit to implement measures to disclose deep fakes that form part of evidently artistic, creative, satirical, fictional or analogous work or programmes in a way 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.

In this context, Signatories will:

a) use an icon or equivalent label following the design specifications in Measure 1.1. and place it in a manner appropriate to the type of artistic, creative, satirical, fictional or analogous work and to the context in which it is presented.

b) ensure such disclosure and placement are clear, distinguishable, and accessible to all natural persons, and provided at the latest at the time of first exposure to the content containing the deep fake (e.g., in the accompanying notes or description provided to users, at the beginning/ end credits etc.).

c) ensure the icon or equivalent label is perceivable for a sufficient duration to be easily noticed under normal viewing or exposure conditions to ensure perceivability.

Where deep fake content is made available in a **digital or interactive** manner (e.g., on websites, apps or other user interfaces), the icon or equivalent label may 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 Signatory. Signatories commit to ensure that any such contextual disclosure solution is perceivable by the end-user without the need to perform dedicated actions or additional engagement. Examples of contextual disclosure include disclosures as part of the user interface (e.g., as part of websites or application interfaces of smart glasses or other digital devices), short notes, a non-obtrusive icon or label that by clicking or hovering over provides more information, disclaimers appearing adjacent to the content.

Where content is made available in a **non-digital or non-interactive manner** (e.g., exhibitions, art galleries, cinemas, festivals or similar contexts, audio or video on a physical carrier), disclosures may be provided at the online or physical point of entry or sale, as part of the introductory or accompanying information (e.g. exhibition leaflet or entrance ticket), or information provided via a physical carrier (e.g. packaging).

Commitment 4: Human Review and Editorial Control for Published Text

Signatories who are media service providers within the meaning of Art. 2(2) of Regulation (EU) 2024/1083 and already subject to editorial standards and to regulatory, co-regulatory, or self-regulatory frameworks, as applicable to those media service providers under EU and national media law and frameworks, may rely on the exception to the disclosure obligation in Article 50(4), subparagraph 2, AI Act by applying their existing review and editorial procedures and established professional standards, as applicable, to comply with this Commitment.

All other Signatories, including those without such review or editorial procedures, commit to establish, adapt, or maintain appropriate policies for human review or editorial control prior to publication and that a natural or legal person holds editorial responsibility for the publication. These internal policies may rely on existing processes, will be proportionate to the deployer's size and resources, and will include at least the following elements:

a) The identification of the natural or legal person with editorial responsibility (name, role and contact details);

b) 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 published text. This does not entail having to document individual instances of human review or editorial control over individual text publications.

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

Signatories may record additional information on the nature of the review or the type of involvement of the AI system in the published text.

The implementation of this Commitment shall in no way affect media freedom, editorial independence and protection of journalistic source information.

Annex

Annex 1: Publicly available EU icons

This Annex contains the EU icons that may be used by deployers to comply with Measure 1.1. Three main icons are provided. One icon is intended to disclose fully AI-generated deep fakes or published text (Figure 1), a second icon is intended to disclose AI-manipulated/partially modified deep fakes or published text (Figure 2), and a third basic icon is included to enable deployers to supplement it with an interactive layer with further information or an alternative textual label, where appropriate (Figure 3).

The design and development of the EU icon have been subject to empirical user-testing conducted across several Member States, which evaluated the perception of authenticity, noticeability and recognition, understandability and trust. The user-tests included various icon designs and demonstrated that while all tested designs achieved high noticeability and reliability, the variants that include a clear textual label (i.e. "modified") performed significantly better in terms of noticeability and clarity for end-users. Respondents participating in the survey indicated that explicit text on AI-generation or manipulation reduces the ambiguity regarding the nature of the content. Consequently, based on this user feedback, the Code includes an icon for fully AI-generated content ("AI + GENERATED") and one for partially manipulated content ("AI + MODIFIED") to increase transparency for the persons exposed to the content and fulfil the disclosure objectives of the AI Act. The findings and methodology of the empirical study will be made public.

These icons are publicly available for Signatories to use in line with the placement specifications, without the need for attribution to the Commission or the AI Office.

The task force will explore the integration of an interactive second layer with the aim to enrich the above static icons with provenance data and information on what has been modified. The development of this interactive layer will be in alignment with the technical capabilities of deployers and other actors distributing and/or verifying content. An-audio-only EU icon will also be developed.

Figure 1. The EU icon developed by the AI Office for fully AI-generated content in several variations against a background (black, white, black transparent, white transparent).

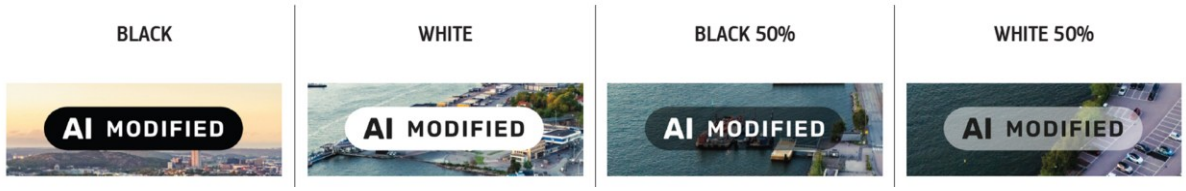

Figure 2. The EU icon developed by the AI Office for partially AI-modified content in several variations against a background (black, white, black transparent, white transparent).

Figure 3. The basic EU icon developed by the AI Office in several variations against a background (black, white, black transparent, white transparent).