

Code de bonnes pratiques sur la transparence des contenus générés par IA

Traduction française non officielle

Document original publié par la Commission européenne (Bureau de l'IA) le 10 juin 2026, dans le cadre de l'article 50 du règlement sur l'IA (AI Act).

Avertissement. Traduction de courtoisie réalisée par TransparIA. Seule la version anglaise publiée par la Commission européenne fait foi. La Commission n'est pas responsable de cette traduction. Source : digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content

Table des matières

Section 1 : Marquage et détection des contenus générés et manipulés par IA applicables aux fournisseurs de systèmes d'IA générative (article 50, paragraphes 2 et 5, du règlement sur l'IA (AI Act))

Objectifs

Considérants

Engagements

Engagement 1 : Marquage du contenu généré ou manipulé par IA

Mesure 1.1 : Techniques de marquage lisible par machine

Sous-mesure 1.1.1 : Métadonnées signées numériquement

Sous-mesure 1.1.2 : Filigrane numérique imperceptible

Sous-mesure 1.1.3 : Empreinte numérique (« fingerprinting ») ou journalisation (facultatif)

Mesure 1.2 : Non-suppression des marquages

Mesure 1.3 : Transparence des informations de provenance (facultatif)

Mesure 1.4 : Fonctionnalité pour les marquages perceptibles (facultatif)

Engagement 2 : Détection des marquages de contenu généré ou manipulé par IA

Mesure 2.1 : Mécanismes de détection des marquages

Sous-mesure 2.1.1. Mise à disposition de la solution de détection

Sous-mesure 2.1.2 : Accès à la solution de détection

Sous-mesure 2.1.3 : Protection des données à caractère personnel, vie privée et sécurité

Sous-mesure 2.1.4 : Retrait d'une solution de détection

Mesure 2.2 : Mécanisme de détection forensique (optionnel)

Mesure 2.3 : Divulcation claire et accessible des résultats de détection

Mesure 2.4 : Soutenir la littératie en matière de marquage de l'IA et de solutions de détection (optionnel)

Engagement 3 : Mesures pour répondre aux exigences en matière de solutions de marquage et de détection

Mesure 3.1 : Efficacité

Mesure 3.2 : Fiabilité

Mesure 3.3 : Robustesse

Mesure 3.4 : Interopérabilité

Mesure 3.5 : Faire progresser l'état de l'art (facultatif)

Engagement 4 : Tests, vérification et conformité

Mesure 4.1 : Processus de conformité

Mesure 4.2 : Tests, vérification et surveillance

Mesure 4.3 : Formation

Mesure 4.4 : Coopération avec les autorités de surveillance du marché

Glossaire

Section 2 : Étiquetage des hypertrucages (« deepfakes ») et des textes publiés générés et manipulés par IA applicable aux déployeurs de systèmes d'IA (Article 50(4) et (5) du règlement sur l'IA (AI Act))

Objectifs

Considérants

Engagements

Engagement 1 : Divulcation des hypertrucages (« deepfakes ») et des textes publiés

Mesure 1.1 : Spécifications de conception

Mesure 1.2 : Spécifications de placement

Sous-mesure 1.2.1 : Principes généraux relatifs aux spécifications de placement

Sous-mesure 1.2.2 : Spécifications de placement lorsqu'une divulgation visuelle est possible

Sous-mesure 1.2.3 : Spécifications de placement lorsqu'une divulgation visuelle n'est pas possible

Mesure 1.3 : Participation volontaire à un groupe de travail au titre du Code

Engagement 2 : Processus internes

Mesure 2.1 : Processus interne de conformité

Mesure 2.2 : Sensibilisation et maîtrise

Mesure 2.3 : Réexamen, retours d'information et coopération avec les autorités

Engagement 3 : Divulgation pour les oeuvres artistiques, créatives et similaires

Engagement 4 : Réexamen humain et contrôle éditorial pour les textes publiés

Annexe

Annexe 1 : Icônes de l'UE accessibles au public

Section 1 : Marquage et détection des contenus générés et manipulés par IA applicables aux fournisseurs de systèmes d'IA générative (article 50, paragraphes 2 et 5, du règlement sur l'IA (AI Act))

Objectifs

L'objectif primordial du présent code de bonnes pratiques (« Code ») est d'améliorer le fonctionnement du marché intérieur, de promouvoir l'adoption d'une intelligence artificielle (« IA ») axée sur l'humain et digne de confiance, et de soutenir l'innovation conformément à l'article 1er, paragraphe 1, du règlement sur l'IA (AI Act), tout en garantissant un niveau élevé de protection de la santé, de la sécurité et des droits fondamentaux consacrés par la Charte, y compris la démocratie, l'État de droit et la protection de l'environnement, contre les effets néfastes de l'IA dans l'Union.

Pour atteindre cet objectif primordial, les objectifs spécifiques de la présente Section du Code sont les suivants :

- a. servir de document d'orientation pour démontrer le respect des obligations prévues à l'article 50, paragraphes 2 et 5, du règlement sur l'IA (AI Act), tout en reconnaissant que l'adhésion au Code ne constitue pas une preuve concluante du respect de ces obligations ;
- b. aider les fournisseurs de systèmes d'IA générant des contenus de synthèse audio, image, vidéo ou texte (« systèmes d'IA générative ») à se conformer à leurs obligations au titre de l'article 50, paragraphes 2 et 5, du règlement sur l'IA (AI Act) ; et
- c. permettre aux autorités de surveillance du marché compétentes d'évaluer ce respect à l'égard des fournisseurs qui s'appuient sur le Code pour démontrer le respect de ces obligations au titre de l'article 50, paragraphes 2 et 5, du règlement sur l'IA (AI Act), d'une manière cohérente, prévisible et uniforme dans toute l'Union.

Considérants

Considérant ce qui suit :

- a. **Confiance dans l'écosystème de l'information** : Les Signataires reconnaissent que les systèmes d'IA peuvent générer et manipuler de grandes quantités de contenu de synthèse et qu'il devient de plus en plus difficile pour les humains de distinguer ce contenu du contenu rédigé par des humains. De telles évolutions ont une incidence sur l'intégrité de l'écosystème de l'information ainsi que sur la confiance que les personnes lui accordent, et soulèvent de nouveaux risques de désinformation, de manipulation à grande échelle, de fraude, d'usurpation d'identité et de tromperie des consommateurs. Les Signataires reconnaissent que la transparence est fondamentale pour favoriser la confiance dans l'écosystème et son intégrité, et pour garantir que les systèmes d'IA demeurent fiables et dignes de confiance.

- b. **Solutions techniques de marquage** : Bien qu'au moment de la publication du présent Code plusieurs techniques soient aisément disponibles pour marquer le contenu audio, image, vidéo ou texte généré ou manipulé par IA, les Signataires reconnaissent que, dans la plupart des cas, en particulier pour le contenu susceptible d'être diffusé en ligne, aucune technique de marquage unique ne suffit à satisfaire aux quatre exigences de l'article 50, paragraphe 2, du règlement sur l'IA (AI Act), à savoir l'efficacité, l'interopérabilité, la robustesse et la fiabilité. Dans de tels cas, les Signataires reconnaissent que, comme cela est admis dans les rapports scientifiques et conformément à l'état de l'art, seule une combinaison appropriée de techniques de marquage et de mécanismes de détection associés peut permettre de satisfaire à ces exigences de manière holistique, dans la mesure où cela est techniquement réalisable, en tenant compte des spécificités et des limites du contenu, du coût de mise en œuvre et de l'existence de normes pertinentes. Le Code promeut l'innovation et la pérennité dans un paysage technologique en évolution rapide en permettant aux Signataires de s'appuyer sur des techniques alternatives, ou même sur une technique unique, pour autant qu'ils puissent prouver le respect des quatre exigences de l'article 50, paragraphe 2, du règlement sur l'IA (AI Act) et de la présente Section du Code, sur la base de résultats vérifiables et de points de référence communs, lorsqu'ils sont disponibles.
- c. **Coopération tout au long de la chaîne de valeur** : Les Signataires reconnaissent que des solutions techniques de marquage et de détection efficaces, interopérables, robustes et fiables nécessitent un investissement en temps et en ressources. Les Signataires reconnaissent donc la nécessité de modalités pratiques visant à rendre les solutions de détection accessibles, le cas échéant, et à faciliter la coopération avec d'autres acteurs tout au long de la chaîne de valeur afin de permettre au public d'identifier efficacement le contenu généré ou manipulé par IA. Les Signataires qui sont fournisseurs de modèles d'IA générative jouent un rôle important dans cette chaîne de valeur et sont donc encouragés à faciliter la mise en conformité des fournisseurs en aval de systèmes d'IA générative construits sur ces modèles.
- d. **Faire progresser l'innovation dans les techniques de marquage et les mécanismes de détection** : Les Signataires reconnaissent que la détermination des méthodes techniques de marquage et de détection les plus efficaces, robustes, fiables et interopérables demeure un défi en constante évolution. Les Signataires reconnaissent que la présente Section du Code promeut l'innovation de diverses manières et encourage à faire progresser l'état de l'art en matière de techniques de transparence de l'IA ainsi que des processus et mesures connexes. Par exemple, le Code permet (i) des techniques innovantes pour la transparence de l'IA, (ii) de nouveaux moyens de faciliter l'interopérabilité entre différentes techniques de marquage et les mécanismes de détection associés, et (iii) le développement de points de référence pour évaluer les performances des solutions techniques mises en œuvre par les fournisseurs de systèmes d'IA générative. Les Signataires sont donc encouragés à investir du temps et des ressources, le cas échéant, pour contribuer à l'avancement de l'état de l'art.

- e. **Coopération avec d'autres parties prenantes** : Les Signataires reconnaissent les avantages d'une efficacité collaborative, par exemple par le partage de méthodes et/ou d'infrastructures et le recours à des normes ouvertes et à des techniques de marquage mises en œuvre au niveau du modèle ou fournies par d'autres tiers. Les Signataires reconnaissent en outre l'importance de permettre aux tiers pertinents et aux utilisateurs finaux de détecter le contenu marqué, et d'associer des représentants experts ou profanes de la société civile, du monde universitaire et d'autres parties prenantes pertinentes à la compréhension des solutions techniques pour la transparence de l'IA. Les Signataires reconnaissent qu'une telle coopération peut impliquer de fournir le soutien nécessaire à la bonne intégration des modèles tiers et des solutions de marquage dans le processus d'inférence du système d'IA, et de conclure des accords de partage d'informations pertinentes pour les solutions techniques, tout en garantissant une protection proportionnée des informations sensibles et le respect du droit de l'Union applicable. Les Signataires reconnaissent également que le retour d'expérience des utilisateurs finaux sur leur mise en œuvre des techniques de marquage et de détection constitue une source d'information précieuse en vue d'en améliorer continuellement la mise en œuvre. Les Signataires reconnaissent en outre l'importance de coopérer avec les autorités de surveillance du marché et de favoriser la collaboration entre les fournisseurs de systèmes et de modèles d'IA générative, les déployeurs et les autres utilisateurs de systèmes d'IA générative, les plateformes en ligne, les moteurs de recherche, les chercheurs, la société civile et les organismes de réglementation, afin de relever les défis et de saisir les opportunités émergentes en matière de transparence des contenus générés ou manipulés par IA.
- f. **Promouvoir la normalisation** : Les Signataires reconnaissent la valeur des normes ouvertes pour faciliter l'interopérabilité et réduire le coût des techniques de marquage et des mécanismes de détection associés pour la transparence de l'IA. Ils reconnaissent que des efforts supplémentaires seront nécessaires pour que de telles normes émergent des organisations internationales et européennes de normalisation, compte tenu des défis de mise en œuvre et du domaine en évolution rapide. Les Signataires sont encouragés à participer aux activités de normalisation pertinentes afin d'accélérer la consolidation des techniques de marquage et des mécanismes de détection associés ainsi que leur interopérabilité. En particulier, les Signataires reconnaissent que les normes de marquage par métadonnées doivent être davantage élaborées pour fournir des informations plus riches en faveur de la transparence de l'IA, tout en reconnaissant la nécessité de normes pertinentes en matière de filigrane numérique, tant au niveau de la technologie que des interfaces.
- g. **Proportionnalité pour les petites et moyennes entreprises (« PME ») et les petites entreprises à moyenne capitalisation (« PEMC »)** : Pour tenir compte des différences entre les fournisseurs de systèmes d'IA générative quant à leur taille et à leur capacité, des modalités de mise en conformité simplifiées pour les PME et les PEMC, y compris les jeunes pousses, devraient être possibles, de manière proportionnée.

Engagements

La présente Section du Code s'applique aux Signataires qui sont fournisseurs de systèmes d'IA générative relevant de l'article 2 et de l'article 50, paragraphe 2, du règlement sur l'IA (AI Act), en tenant compte de la portée des obligations légales et des exceptions pertinentes, le cas échéant.

Sans préjudice de la responsabilité première des fournisseurs de systèmes d'IA générative au titre de l'article 50, paragraphe 2, du règlement sur l'IA (AI Act), les fournisseurs de modèles d'IA générative qui sont mis sur le marché de l'Union indépendamment des systèmes d'IA, ainsi que les fournisseurs tiers de solutions de marquage et de détection, peuvent également, sur une base volontaire, adhérer au Code afin de démontrer que leur marquage et

détection respectent les exigences de l'article 50, paragraphes 2 et 5, du règlement sur l'IA (AI Act) et de la présente Section du Code, en vue de faciliter la mise en conformité des fournisseurs en aval de systèmes d'IA construits sur ces modèles et/ou utilisant les solutions de marquage et de détection de tiers. Dans de tels cas, les engagements et les mesures de la présente Section du Code s'entendent comme s'appliquant au modèle d'IA générative et/ou à la solution de marquage et de détection du Signataire, selon le cas.

Dans la présente Section, les termes « will », « encouraged » et « may » doivent être interprétés de la manière suivante :

1. « will » est utilisé pour les mesures obligatoires du Code qui doivent être respectées pour que le Signataire soit en conformité avec l'article 50, paragraphes 2 et 5, du règlement sur l'IA (AI Act), et dont le respect sera contrôlé par les autorités de surveillance du marché compétentes ;
2. « encouraged » est utilisé pour les mesures facultatives du Code qui ne sont pas juridiquement requises et sont purement volontaires, mais qui sont néanmoins recommandées ;
3. « may » est utilisé pour les mesures facultatives du Code qui ne sont pas juridiquement requises et sont purement volontaires ou qui offrent aux fournisseurs une certaine souplesse quant à la manière de mettre en oeuvre la mesure ou l'engagement concerné.

Engagement 1 : Marquage du contenu généré ou manipulé par IA

TEXTE JURIDIQUE : article 50, paragraphe 2, et considérants 133 et 135 du règlement sur l'IA (AI Act)

2. Les fournisseurs de systèmes d'IA, y compris de systèmes d'IA à usage général, générant des contenus de synthèse de type audio, image, vidéo ou texte, veillent à ce que les sorties du système d'IA soient marquées dans un format lisible par machine et détectables comme ayant été générées ou manipulées par une machine. Les fournisseurs veillent à ce que leurs solutions techniques soient efficaces,interopérables, robustes et fiables dans la mesure où cela est techniquement possible, compte tenu des spécificités et des limites des différents types de contenus, des coûts de mise en oeuvre et de l'état de l'art généralement reconnu, tel qu'il peut être reflété dans les normes techniques pertinentes.

Afin de remplir leur obligation au titre de l'article 50, paragraphe 2, du règlement sur l'IA (AI Act) de marquer de manière lisible par machine les sorties des systèmes d'IA générative, y compris des systèmes d'IA à usage général, les Signataires s'engagent à mettre en oeuvre une solution de marquage à l'égard des contenus audio, image, vidéo ou texte, ou de toute combinaison de ceux-ci, générés ou manipulés par le ou les systèmes d'IA qu'ils mettent sur le marché ou mettent en service dans l'Union.

Afin de remplir cet Engagement, les Signataires s'engagent à mettre en oeuvre les Mesures suivantes, selon ce qui est applicable au(x) type(s) de contenu généré(s) ou manipulé(s) par leur(s) système(s) d'IA.

Mesure 1.1 : Techniques de marquage lisible par machine

Les Signataires mettent en oeuvre une solution de marquage qui consiste en au moins une technique de marquage lisible par machine, telle que détaillée ci-dessous, qui, conjointement avec leur mécanisme de détection respectif, atteint le niveau d'efficacité, de fiabilité, de robustesse et d'interopérabilité requis par l'article 50, paragraphe 2, du règlement sur l'IA (AI Act) et par l'Engagement 3 du Code.

Tant qu'aucune technique de marquage unique ne peut, dans l'état de l'art, assurer à elle seule la conformité aux quatre exigences de l'article 50, paragraphe 2, du règlement sur l'IA (AI Act) d'efficacité, d'interopérabilité, de robustesse et de fiabilité pour l'audio, les images, la vidéo et le texte conteneurisé, en particulier pour les contenus susceptibles d'être diffusés en ligne, les Signataires mettent en oeuvre une approche de marquage multicouche afin de garantir que les sorties de leurs systèmes d'IA générative soient marquées avec au moins deux couches de marquage lisible par machine, comme précisé dans les Sous-mesures 1.1.1 et 1.1.2 ci-dessous.

Pour permettre une mise en oeuvre proportionnée des exigences ci-dessus, une couche unique de marquage sera considérée comme suffisante dans des cas spécifiques où un système d'IA générative est intégré dans des produits physiques capables de générer des sorties de synthèse dans un environnement techniquement contrôlé et fermé, à caractère principalement instructif. Cela s'applique dans la mesure où des mesures techniques efficaces sont intégrées dans le produit pour empêcher la sortie de quitter l'environnement du produit, par exemple en étant captée ou exportée et diffusée à l'extérieur. En outre, étant donné que le texte libre ne peut transporter de métadonnées, une couche unique de marquage telle que décrite dans la Sous-mesure 1.1.2 est considérée comme suffisante pour se conformer aux exigences de l'article 50, paragraphe 2, du règlement sur l'IA (AI Act) pour ce type spécifique de contenu.

À l'avenir, les Signataires pourront également démontrer leur conformité à l'article 50, paragraphe 2, du règlement sur l'IA (AI Act) par le recours à une ou des techniques de marquage alternatives (éventuellement une technique de marquage unique), à condition qu'ils puissent prouver aux autorités de surveillance du marché compétentes, sur la base de méthodes d'évaluation des performances et de référentiels reconnus, que leur(s) technique(s) atteint/atteignent au moins le même niveau de robustesse, de fiabilité, d'efficacité et d'interopérabilité, sinon supérieur, que celui requis par l'article 50, paragraphe 2, du règlement sur l'IA (AI Act) et par l'Engagement 3 du Code. Dans l'attente de l'émergence de méthodes d'évaluation des performances et de référentiels reconnus, et comme précisé dans la Mesure 4.2, les Signataires peuvent démontrer leur conformité au moyen de tests internes documentés au regard des exigences de l'Engagement 3, sous réserve d'un examen par les autorités de surveillance du marché compétentes.

Les techniques de marquage peuvent être mises en oeuvre à différents stades de la chaîne de valeur (par exemple, par le fournisseur du système d'IA lui-même ou par un fournisseur de modèle en amont) et peuvent également être fournies par des tiers (en particulier, des fournisseurs de technologies spécialisés dans les techniques de marquage et les mécanismes de détection pour la transparence de l'IA). Les Signataires peuvent s'appuyer sur ces solutions techniques de tiers sans préjudice de la responsabilité propre des Signataires, au titre du règlement sur l'IA (AI Act) et du Code, de veiller à ce que les sorties de leurs systèmes d'IA soient marquées de manière appropriée et conforme.

Sous-mesure 1.1.1 : Métadonnées signées numériquement

Si un contenu est généré, manipulé ou exporté dans un format de données qui prend en charge l'attachement de métadonnées (par exemple, un audio, une image, une vidéo ou un texte conteneurisé), les Signataires enregistrent dans les métadonnées l'information indiquant si le contenu est généré ou manipulé par IA.

Toutes les informations enregistrées seront signées numériquement et horodatées (sur les systèmes où l'information temporelle est disponible) de manière sécurisée et à l'épreuve de toute altération. Les Signataires adoptent des moyens visant à garantir l'usage sécurisé des certificats de signature et la confidentialité des clés privées associées, sauf dans les circonstances où le contexte de déploiement ne permet pas le traitement ou la fourniture sécurisés de la clé, par exemple dans un scénario de déploiement local.

Les Signataires sont encouragés à mettre en oeuvre des métadonnées plus riches conformément à la Mesure 1.3, sans inclure d'informations sensibles sur le plan de la vie privée ou sensibles sur le plan commercial. Dans les cas où de telles informations sont strictement nécessaires ou insérées à la demande d'un utilisateur final, elles seront conformes au droit de l'Union applicable en matière de protection des données et il est encouragé de les placer dans un emplacement réservé de métadonnées distinct des métadonnées de transparence de l'IA.

Sous-mesure 1.1.2 : Filigrane numérique imperceptible

Les Signataires veillent à ce que le contenu généré ou manipulé par IA soit marqué d'un filigrane numérique imperceptible, à l'exception des textes très courts. Pour le texte libre de plus de 200 jetons, le filigranage doit néanmoins être appliqué, même s'il peut présenter une fiabilité moindre par rapport à celle du filigranage d'un texte très long. Pour compenser la fiabilité potentiellement moindre du filigranage pour le texte libre, l'accès à la solution de détection correspondante peut être restreint à des utilisateurs experts vérifiés, comme détaillé dans la Sous-mesure 2.1.2.

Le filigrane numérique imperceptible sera intégré au sein du contenu d'une manière qui rende difficile sa séparation du contenu. Le filigrane numérique est destiné à servir de mécanisme robuste pour compléter les métadonnées signées numériquement au titre de la Sous-mesure 1.1.1.

L'état de l'art offre plusieurs stratégies pour intégrer des filigranes numériques dans un contenu généré ou manipulé par IA, par exemple en appliquant la technique de filigranage une fois le contenu généré ou manipulé (c'est-à-dire le « filigranage post-hoc ») ou en faisant introduire le filigrane numérique pendant l'opération d'inférence du système d'IA générative (c'est-à-dire le « filigranage du modèle »). Les Signataires qui fournissent des modèles d'IA générative sont encouragés à mettre en oeuvre le filigranage au niveau du modèle et à permettre l'intégration fluide du filigranage dans le processus d'inférence du système d'IA, afin de faciliter la mise en conformité des fournisseurs en aval de systèmes d'IA générative construits sur ces modèles, en particulier d'une manière qui aide ces fournisseurs en aval à satisfaire aux exigences de qualité précisées à l'article 50, paragraphe 2, du règlement sur l'IA (AI Act) et dans l'Engagement 3 et qui leur permette de démontrer leur conformité conformément à l'Engagement 4.

Sous-mesure 1.1.3 : Empreinte numérique (« fingerprinting ») ou journalisation (facultatif)

Lorsque cela est approprié et compte tenu des compromis potentiels liés à la vie privée et à la sécurité, ainsi que des défis de passage à l'échelle et des coûts, les Signataires peuvent mettre en oeuvre, en tant que mesure supplémentaire facultative, des solutions d'empreinte numérique (« fingerprinting ») ou de journalisation pour les contenus générés ou manipulés par IA qui permettent de vérifier si un contenu a été généré ou manipulé par leur système d'IA. Par exemple, la journalisation directe peut être appropriée pour le contenu texte, tandis que les approches d'empreinte numérique (« fingerprinting ») peuvent être préférables pour le contenu audio et visuel.

L'utilisation de telles techniques peut constituer une solution supplémentaire pour satisfaire aux exigences de qualité précisées à l'article 50, paragraphe 2, du règlement sur l'IA (AI Act) et dans l'Engagement 3, ou pour contribuer à réduire le coût du mécanisme de détection décrit dans l'Engagement 2. Toutefois, le recours à l'empreinte numérique (« fingerprinting ») ou à la journalisation seules n'est pas considéré comme suffisant pour satisfaire aux exigences de qualité précisées à l'article 50, paragraphe 2, du règlement sur l'IA (AI Act) et dans l'Engagement 3.

Lorsque les Signataires mettent en oeuvre l'empreinte numérique (« fingerprinting ») ou la journalisation, ils veillent à ce qu'elle soit limitée aux données de sortie et qu'elle soit mise en oeuvre de manière sécurisée et respectueuse de la vie privée, dans le respect du droit de l'Union en matière de protection des données et dans le respect de la liberté des médias, de l'indépendance éditoriale et des obligations de protection des sources journalistiques. Les déployeurs et autres utilisateurs des systèmes d'IA générative se verront accorder l'accès aux politiques de journalisation ainsi que le contrôle de ce qui est journalisé et de la manière dont les données sont stockées, consultées et de la durée pendant laquelle elles sont conservées, avec des procédures claires pour des durées de conservation limitées, des contrôles d'accès, la sécurité et d'autres garanties de minimisation des données, ainsi que la suppression sécurisée et immédiate des journaux contenant des données de sortie une fois leur finalité prévue accomplie.

Cette Sous-mesure est limitée à la conception et à la mise en oeuvre propres du système d'IA par le Signataire et ne saurait en aucun cas être interprétée comme créant un engagement général de journaliser, de surveiller ou de conserver les invites pour le contenu généré ou manipulé par IA ou pour les interactions des utilisateurs.

Mesure 1.2 : Non-suppression des marquages

Sans préjudice de l'exigence de robustesse de l'Engagement 3, les Signataires mettent tout en oeuvre pour préserver les marquages de métadonnées sur les données d'entrée et sur le contenu généré ou manipulé par leur système d'IA en appliquant les mesures cumulatives suivantes :

- a. Les Signataires, dans la mesure où cela est techniquement possible et reconnaissable selon des normes ouvertes, conservent les marquages de métadonnées existants, et s'abstiennent de les altérer ou de les supprimer intentionnellement, lorsqu'un tel contenu est utilisé comme entrée puis transformé par leur système d'IA en une sortie. Cela n'affecte pas le traitement de bonne foi et légitime lorsque la modification, la transformation ou le remplacement des métadonnées existantes est nécessaire pour maintenir des informations exactes et fonctionnelles à la suite d'un traitement en aval ou d'un traitement légitime spécifique, par exemple à des fins d'audits de sécurité et de recherche.
- b. Bien que les Signataires reconnaissent que la conformité et l'application en aval ne peuvent être garanties, ils incluent néanmoins dans la politique d'utilisation acceptable, les conditions générales ou la documentation accompagnant leur système d'IA générative une interdiction de la suppression intentionnelle ou de l'altération des marquages de métadonnées par les déployeurs ou tout autre tiers. Les exceptions à une telle interdiction sont les finalités légitimes visées au point a). Pour les systèmes d'IA et les modèles fournis sous licence libre et open source, il suffit que les Signataires alertent les utilisateurs sur cette bonne pratique dans la documentation accompagnant le système d'IA ou le modèle, sans préjudice de la nature libre et open source de la licence.

Les mesures précisées au point a) ci-dessus n'impliquent pas la responsabilité du Signataire à l'égard des marquages de métadonnées de tiers, ni à l'égard de la conformité de tiers aux conditions précisées au point b).

En outre, les Signataires ne mettent pas sur le marché ni ne mettent à disposition sur le marché, et ne font pas la promotion ou la publicité de l'utilisation, d'outils dont la finalité est de contourner les marquages lisibles par machine ajoutés au contenu généré ou manipulé par IA à des fins de transparence.

Les Signataires qui exploitent une plateforme en ligne ou un moteur de recherche, ou qui diffusent par ailleurs des contenus au public, sont encouragés à veiller à ce que la plateforme en ligne ou le moteur de recherche préserve les marquages de métadonnées pour le contenu généré ou manipulé par IA.

Mesure 1.3 : Transparence des informations de provenance (facultatif)

Les Signataires sont encouragés à incorporer des informations plus riches dans les métadonnées conformément à la Mesure 1.1.1, afin de fournir un contexte supplémentaire et de contribuer ainsi à une intégrité et à une confiance accrues dans l'écosystème de l'information. En particulier, les Signataires sont encouragés à prendre en compte les normes pertinentes pour fournir des informations supplémentaires sur l'origine du contenu généré ou manipulé par IA à travers les flux de travail, lorsque cela est techniquement possible.

Les Signataires sont encouragés à ajouter ou à enregistrer des informations de provenance pertinentes au sein du contenu généré ou manipulé par leurs systèmes d'IA, en particulier le nom du système d'IA, la raison sociale du fournisseur d'IA, et un horodatage indiquant à quel moment le contenu a été généré ou manipulé. En outre, les Signataires peuvent également inclure l'identifiant du modèle d'IA sous-jacent et son numéro de version.

Pour le contenu manipulé par IA, les Signataires sont encouragés à enregistrer des informations sur le type d'opérations effectuées par leur système d'IA pour modifier le contenu (par exemple, la suppression d'objet). Il est recommandé d'encoder de multiples opérations distinctes effectuées par le système d'IA dans un marqueur de métadonnées unique afin de limiter la complexité et de réduire la charge.

Mesure 1.4 : Fonctionnalité pour les marquages perceptibles (facultatif)

Afin de faciliter la mise en conformité des déployeurs de systèmes d'IA générative avec leur obligation de divulguer les hypertrucages (« deepfakes ») et le texte publié généré ou manipulé par IA conformément à l'article 50, paragraphe 4, du règlement sur l'IA (AI Act), les Signataires qui sont des fournisseurs de systèmes d'IA générative capables de générer ou de manipuler de tels contenus sont encouragés à fournir une fonctionnalité facultative dans l'interface de leur système et à mettre en oeuvre une option intégrée qui permette aux déployeurs et autres utilisateurs d'appliquer directement, à leur propre discrétion, lors de la génération de la sortie, un label perceptible en cohérence avec les Engagements et Mesures de la Section 2 du Code.

Les Signataires sont également encouragés à suivre des normes d'expérience utilisateur (UX) harmonisées, dans la mesure du possible, d'une manière interopérable avec les systèmes et flux de travail de gestion de contenu normalisés existants des plateformes en ligne et des éditeurs de médias.

Les Signataires sont en outre encouragés à mettre en oeuvre d'autres mesures de soutien pour l'affichage des labels et des métadonnées qui permettent aux déployeurs et aux fournisseurs de plateformes en ligne, de moteurs de recherche et de sites web de mettre en oeuvre des pratiques et des politiques d'affichage adaptées à leurs cas d'usage.

Cette mesure s'entend sans préjudice de la responsabilité des déployeurs, qui restent responsables de la divulgation des hypertrucages (« deepfakes ») et du texte publié généré ou manipulé par IA de manière claire et distinguable conformément à l'article 50, paragraphes 4 et 5, du règlement sur l'IA (AI Act).

Engagement 2 : Détection des marquages de contenu généré ou manipulé par IA

TEXTE JURIDIQUE : Article 50, paragraphes 2 et 5, et considérants 133 et 135 du règlement sur l'IA (AI Act)

2. Les fournisseurs de systèmes d'IA, y compris de systèmes d'IA à usage général, qui génèrent du contenu de synthèse de type audio, image, vidéo ou texte, veillent à ce que les sorties du système d'IA soient [...] détectables comme ayant été générées ou manipulées par des moyens artificiels.
5. Les informations visées aux paragraphes 1 à 4 sont fournies aux personnes physiques concernées de manière claire et reconnaissable au plus tard au moment de la première interaction ou exposition. Ces informations sont conformes aux exigences applicables en matière d'accessibilité.

Afin de remplir leurs obligations au titre de l'article 50, paragraphes 2 et 5, du règlement sur l'IA (AI Act) consistant à veiller à ce que les sorties de leur(s) système(s) d'IA soient détectables comme ayant été générées ou manipulées par IA, les Signataires s'engagent à fournir les moyens permettant la détection des marquages lisibles par machine présents dans des contenus audio, image, vidéo ou texte, ou une combinaison de ceux-ci, en tant que contenus générés ou manipulés par leur système d'IA. Les Signataires s'engagent également à veiller à ce que le résultat de la détection soit fourni aux personnes physiques concernées de manière claire, reconnaissable et accessible.

Afin de remplir cet Engagement, les Signataires s'engagent à mettre en oeuvre les Mesures suivantes, selon ce qui est applicable au(x) type(s) de contenu généré(s) ou manipulé(s) par leur(s) système(s) d'IA.

Mesure 2.1 : Mécanismes de détection des marquages

Les Signataires mettent à disposition une solution de détection, composée d'un ou de plusieurs mécanismes de détection, afin de permettre aux déployeurs, aux utilisateurs de leur système d'IA générative, aux intégrateurs tiers, aux utilisateurs finaux exposés au contenu, et à d'autres parties légitimes (telles que les autorités compétentes, les chercheurs indépendants, la société civile et les organisations de médias) de vérifier si un contenu a été généré ou manipulé par leur système d'IA sur la base de la ou des techniques de marquage mises en oeuvre conformément à l'Engagement 1. Les Signataires veillent à ce que la solution de détection, conjointement avec la solution de marquage associée, atteigne le niveau d'efficacité, d'interopérabilité, de robustesse et de fiabilité requis par l'article 50, paragraphe 2, du règlement sur l'IA (AI Act), par l'Engagement 3 du Code, et par les Sous-mesures suivantes.

Sous-mesure 2.1.1. Mise à disposition de la solution de détection

La solution de détection peut être mise à disposition dans l'Union sous une ou plusieurs des formes suivantes : (i) une spécification publique, idéalement normalisée, permettant à tout tiers de mettre en oeuvre un mécanisme de détection ; (ii) un logiciel (par exemple, un exécutable autonome ou une bibliothèque) ; (iii) un service en nuage (cloud) accessible aux utilisateurs dans l'Union via une API. Lorsqu'ils fournissent une spécification publique au titre du point (i), les Signataires sont encouragés à fournir une implémentation de référence. Lorsqu'ils fournissent un logiciel au titre du point (ii), les Signataires sont encouragés à veiller à ce que la version soit portable sur la plupart des systèmes d'exploitation (y compris, lorsque cela est techniquement faisable, les systèmes de téléphonie mobile) et à ce que les exigences matérielles soient limitées, voire inexistantes, afin de faciliter l'exécution locale du ou des mécanismes de détection sur tout appareil à usage général.

Les Signataires veillent à ce que leur solution de détection inclue un mécanisme de détection pour chaque technique de marquage mise en oeuvre dans leur système d'IA générative conformément à l'Engagement 1. À l'instar des techniques de marquage de la Mesure 1.1, les Signataires peuvent s'appuyer sur des mécanismes de détection partagés ou tiers, par exemple fournis par des fournisseurs de modèles d'IA qui appliquent des marquages au niveau du modèle ou par des fournisseurs de technologies tiers spécialisés dans les techniques de marquage de transparence et les mécanismes de détection. Un tel recours s'entend sans préjudice de la responsabilité propre des Signataires de veiller au respect de l'article 50, paragraphes 2 et 5, du règlement sur l'IA (AI Act) et de la présente Section du Code. Afin de faciliter la mise en conformité des fournisseurs en aval de systèmes d'IA générative, les Signataires qui sont également fournisseurs de modèles d'IA générative sont encouragés à fournir des mécanismes de détection pour le contenu généré ou manipulé par leurs modèles avant la mise sur le marché du modèle.

Les Signataires mettent la solution de détection à disposition gratuitement. Les Signataires comptant moins de 1 000 000 d'utilisateurs mensuels de leur système d'IA générative, dont la solution de détection engendre des coûts opérationnels substantiels (en particulier lorsqu'elle est fournie pour la détection de filigranes numériques sous forme de service en nuage), peuvent facturer des frais pour l'utilisation de leur solution de détection (qui seront raisonnables, équitables et proportionnés aux coûts opérationnels globaux de cette solution) dans les cas où les demandes émanant d'un seul utilisateur dépassent un seuil raisonnable en raison d'un volume important de requêtes. Tous les Signataires fournissent toujours un accès gratuit à leur solution de détection, sans aucune restriction sur le volume de requêtes, aux autorités compétentes de surveillance du marché et autres régulateurs, aux autorités répressives, aux médias, aux vérificateurs de faits (fact-checkers), aux signaleurs de confiance, aux chercheurs indépendants, aux établissements d'enseignement et de recherche, et aux organisations de la société civile.

Les Signataires sont encouragés à collaborer avec les acteurs pertinents au sein de l'écosystème afin de rendre leur solution de détection directement disponible au sein des plateformes de distribution et de

communication.

Sous-mesure 2.1.2 : Accès à la solution de détection

Les Signataires assurent l'accès à leur solution de détection au moyen d'une interface utilisateur adaptée au public d'utilisateurs finaux susceptibles d'être exposés au contenu généré ou manipulé par leur système d'IA. Lorsque le grand public est susceptible d'être exposé au contenu généré ou manipulé par IA, la solution de détection est mise à la disposition du grand public en conséquence. Toutefois, lorsqu'il existe des garde-fous garantissant que seul un nombre limité de personnes physiques sera exposé au contenu généré ou manipulé par IA (par exemple, les systèmes d'IA en contexte professionnel), et lorsqu'il existe des garde-fous efficaces empêchant une diffusion ultérieure du contenu au grand public, l'accès à la solution de détection peut être restreint à ces personnes physiques.

Dans leur solution de détection, les Signataires peuvent restreindre l'accès aux mécanismes de détection associés aux techniques de filigranage pour le texte libre dans la mesure où ceux-ci présentent un niveau inférieur de fiabilité et de robustesse et où ils sont susceptibles de produire des résultats trompeurs ou de faible confiance. Néanmoins, les résultats de tels mécanismes de détection peuvent tout de même fournir des informations précieuses à des utilisateurs finaux experts vérifiés ayant un besoin légitime, y compris les autorités compétentes de surveillance du marché, les autres régulateurs, les autorités répressives, les médias, les vérificateurs de faits (fact-checkers), les signaleurs de confiance, les chercheurs indépendants, les organisations d'enseignement et de recherche, et les organisations de la société civile. Par conséquent, les Signataires veillent à ce que l'accès pour ces utilisateurs finaux experts soit accordé sous réserve de contrôles d'accès et de garde-fous appropriés. Toute restriction d'accès est limitée dans le temps, jusqu'à ce que des mécanismes de détection plus fiables et plus robustes aient émergé et aient été adoptés à mesure que l'état de l'art des mécanismes de détection pour le filigranage de texte libre évolue.

Les Signataires veillent à ce que les résultats de détection de leur solution de détection soient fournis conformément à la Mesure 2.3 ci-dessous.

En outre, les Signataires veillent à ce que le résultat de détection d'un contenu spécifique soumis à la détection puisse être téléchargé sur demande dans un format signé numériquement, comprenant au moins une empreinte (hash) du contenu soumis à la détection, une URL ou un autre identifiant de la solution de détection, et un horodatage.

Sous-mesure 2.1.3 : Protection des données à caractère personnel, vie privée et sécurité

Les Signataires veillent à ce que la détection soit conforme au droit de l'Union en matière de vie privée et de protection des données.

Sans préjudice du droit applicable de l'Union, y compris le droit de l'Union en matière de vie privée et de protection des données, pour toute solution de détection fournie sous forme de logiciel ou de service (en tout ou en partie) qui nécessite le téléversement de contenu, les Signataires veillent à :

- a. Ce que le contenu soumis à la détection soit traité comme sensible du point de vue de la vie privée et traité dans le respect du droit de l'Union en matière de vie privée et de protection des données.
- b. La confidentialité et l'intégrité du contenu, tant en transit qu'au repos.

- c. La minimisation des données dans la fourniture du service de détection, à savoir que le service de détection ne collecte, ne conserve ni ne traite d'aucune autre manière de données à caractère personnel provenant des utilisateurs de la solution de détection, ou du contenu qui lui est soumis, au-delà de ce qui est nécessaire pour fournir les fonctions de détection, sauf lorsque cela est nécessaire et fondé sur une base juridique valable pour la sécurité et la prévention des abus, pour garantir que la solution de détection fonctionne efficacement, pour un retour d'information des utilisateurs sur la base du volontariat (opt-in) ou pour l'amélioration du service en fonction des préférences des utilisateurs. Lorsque cela est techniquement faisable, les Signataires sont encouragés à appliquer des technologies de préservation de la vie privée et à ne requérir que le téléversement d'un dérivé du contenu qui ne donne pas accès à la sémantique du contenu.
- d. Ce que le contenu soumis à la détection soit traité aux seules fins de détecter les marquages et ne soit pas traité à d'autres fins, sauf lorsque cela est nécessaire pour les autres finalités spécifiées au point (c) ci-dessus.
- e. Ce que le contenu ne soit stocké que pour la durée de la détection et soit définitivement supprimé immédiatement après (c'est-à-dire selon une politique de « rétention zéro »), étant entendu que des données minimales (telles que des journaux de trafic) peuvent être conservées pendant une période limitée sur une base juridique valable aux seules fins de garantir la sécurité et de prévenir les abus. En particulier, les Signataires ne conservent pas de copie verbatim du contenu et le suppriment immédiatement après la détection.
- f. Ce que des garde-fous de sécurité pertinents, adaptés au risque, soient en place pour garantir la confidentialité, l'intégrité, la disponibilité et la résilience continues du service de détection.
- g. Le respect de toutes les autres exigences applicables au titre du droit de l'Union en matière de vie privée et de protection des données, y compris les exigences régissant les transferts de données à caractère personnel vers des pays tiers au titre du droit de l'Union en matière de protection des données.
- h. Ce que les meilleures pratiques en matière de cybersécurité soient mises en oeuvre pour protéger l'accès à la solution de détection dans les cas où un accès non contrôlé à la solution de détection pourrait compromettre la sécurité des techniques de marquage et des mécanismes de détection sous-jacents.

Sous-mesure 2.1.4 : Retrait d'une solution de détection

Les Signataires sont en droit de retirer une solution de détection pour autant qu'elle soit remplacée par une solution de détection alternative. Les Signataires veillent à ce que toute solution de détection alternative de ce type offre des capacités de détection identiques ou supérieures et soit rétrocompatible, permettant la détection de contenu préalablement marqué.

Lorsqu'un Signataire a développé sa propre solution de détection et cesse de fournir cette solution sans remplacement pour quelque raison que ce soit (par exemple, cessation d'activité), il met cette solution à la disposition des autorités compétentes de surveillance du marché, sous réserve de garde-fous de confidentialité et de sécurité, afin de leur permettre de détecter, si nécessaire, le contenu hérité (legacy) généré ou manipulé par leur système d'IA.

Mesure 2.2 : Mécanisme de détection forensique (optionnel)

Pour compléter les techniques de marquage spécifiées dans l'Engagement 1 et sous réserve de leurs capacités, les Signataires peuvent inclure dans le cadre de leur solution de détection, en tant que mesure optionnelle supplémentaire, un mécanisme de détection forensique pour détecter le contenu généré ou manipulé par leur système d'IA ou modèle sous-jacent dont le marquage a été supprimé. Lorsque les Signataires mettent en oeuvre un tel mécanisme de détection forensique, ils veillent à ce que les exigences de vie privée et de sécurité spécifiées à la Sous-mesure 2.1.3 soient satisfaites et à ce que la divulgation du résultat de détection soit claire et accessible comme spécifié à la Mesure 2.3, et déploient leurs meilleurs efforts pour veiller à ce que leur mécanisme de détection forensique soit aligné sur l'état de l'art comme spécifié à l'Engagement 3.

Au moment de la publication du Code, les mécanismes de détection forensique n'étaient pas jugés suffisamment matures pour se conformer aux exigences de qualité énoncées à l'article 50, paragraphe 2, du règlement sur l'IA (AI Act) et à l'Engagement 3. Tant qu'aucun meilleur mécanisme de détection forensique n'a émergé au titre de l'état de l'art, les Signataires qui mettent en oeuvre des mécanismes de détection forensique en tant que mesure optionnelle peuvent restreindre l'accès aux résultats de détection fournis par de tels mécanismes de détection forensique à des utilisateurs experts vérifiés, conformément à l'exception relative aux mécanismes de détection pour le filigrane de texte libre décrite à la Sous-mesure 2.1.2.

Mesure 2.3 : Divulgation claire et accessible des résultats de détection

Afin de remplir leur obligation au titre de l'article 50, paragraphe 5, du règlement sur l'IA (AI Act), les Signataires veillent à ce que les résultats de détection fournis par leur solution de détection soient présentés d'une manière claire et facilement compréhensible pour les personnes physiques exposées au contenu généré ou manipulé par leur système d'IA et qui souhaitent en vérifier l'origine.

Les Signataires veillent à ce que les résultats de détection indiquent s'ils sont fondés sur un marquage par métadonnées, un marquage par filigrane numérique, une détection forensique ou d'autres techniques, dans la mesure où cela est techniquement faisable.

Lorsque des informations supplémentaires sont disponibles dans le filigrane numérique ou dans les métadonnées, les Signataires incorporent ces informations dans les résultats de détection. Les Signataires peuvent également inclure des informations complémentaires dans les résultats de détection, telles que la proportion du contenu qui a été générée ou manipulée par leur système d'IA et/ou la localisation des modifications de l'IA dans le contenu. Pour les résultats de détection détaillés, les Signataires sont encouragés à envisager un accès par paliers afin d'éviter de submerger les utilisateurs finaux d'informations potentiellement déroutantes, en particulier lorsque l'utilisateur final de la solution de détection est un membre du grand public dépourvu de connaissances spécialisées.

Le cas échéant, les Signataires veillent à ce que les interfaces humaines utilisées pour présenter les résultats de détection soient accessibles aux personnes en situation de handicap, dans le respect des exigences applicables en matière d'accessibilité au titre du droit de l'Union, en particulier celles énoncées dans la directive (UE) 2019/882 (l'acte législatif européen sur l'accessibilité) et la directive (UE) 2016/2102 (la directive sur l'accessibilité du web). Les Signataires sont encouragés à mettre en oeuvre toute norme pertinente disponible, y compris, mais sans s'y limiter, la norme harmonisée ETSI EN 301 549 « Exigences d'accessibilité pour les produits et services TIC », et les WCAG 2.1 niveau AA « Règles pour l'accessibilité des contenus web ».

Mesure 2.4 : Soutenir la littératie en matière de marquage de l'IA et de solutions de détection (optionnel)

Les Signataires sont encouragés à veiller à ce que la documentation et autres informations pertinentes (à l'exclusion des informations commerciales confidentielles et des secrets d'affaires) soient fournies aux déployeurs et autres utilisateurs experts afin de les aider à prendre des décisions éclairées quant au marquage et aux solutions de détection associées qu'ils peuvent utiliser, notamment en les aidant à comprendre comment accéder aux solutions de détection, comment effectuer des détections et comment interpréter les résultats de détection.

Outre les supports destinés aux déployeurs, les Signataires sont également encouragés à veiller à ce que des ressources de littératie destinées aux utilisateurs finaux soient fournies, selon ce qui est approprié, et calibrées sur les besoins des profanes et des utilisateurs finaux, y compris lorsque les systèmes d'IA s'adressent à des populations dont la littératie en matière d'IA est plus faible ou dans des contextes sensibles (par exemple, contextes éducatifs ou contextes impliquant des utilisateurs jeunes ou âgés). Ces ressources peuvent être développées par les Signataires eux-mêmes ou créées conjointement dans le cadre d'efforts impliquant d'autres fournisseurs ou par des organisations ou initiatives auxquelles ils appartiennent.

Cette mesure est mise en oeuvre de manière proportionnée, en tenant compte du niveau de sensibilisation des déployeurs et autres utilisateurs du système d'IA générative et des utilisateurs finaux exposés au contenu, ainsi que de la taille et des ressources du fournisseur du système d'IA, en particulier en ce qui concerne les PME et les PMC.

Les Signataires sont encouragés à collaborer avec le monde universitaire, la société civile, les médias et autres organisations pertinentes afin de promouvoir la littératie et la sensibilisation concernant la transparence de l'IA, et à soutenir les initiatives au niveau de l'UE visant à favoriser une compréhension cohérente de la provenance et de la détection dans l'ensemble des États membres.

Engagement 3 : Mesures pour répondre aux exigences en matière de solutions de marquage et de détection

TEXTE JURIDIQUE : article 50, paragraphe 2, et considérant 133 du règlement sur l'IA (AI Act)

2. [...] Les fournisseurs veillent à ce que leurs solutions techniques soient efficaces, interopérables, robustes et fiables dans la mesure où cela est techniquement réalisable, compte tenu des spécificités et des limites des différents types de contenu, des coûts de mise en œuvre et de l'état de l'art généralement reconnu, tel qu'il peut être reflété dans les normes techniques pertinentes.

Afin de remplir leur obligation, au titre de l'article 50, paragraphe 2, du règlement sur l'IA (AI Act), de veiller à ce que les solutions techniques employées pour le marquage et la détection du contenu généré ou manipulé par IA soient efficaces, interopérables, robustes et fiables, dans la mesure où cela est techniquement réalisable et compte tenu des spécificités et des limites des différents types de contenu, des coûts de mise en œuvre et de l'état de l'art généralement reconnu, les Signataires s'engagent à veiller à ce que leurs solutions de marquage et de détection respectent les exigences de qualité décrites dans les mesures suivantes, de manière globale pour l'ensemble des techniques de marquage employées plutôt que pour chaque technique individuellement. Les Signataires s'engagent à évaluer et à démontrer la conformité aux exigences conformément aux processus d'essai, de vérification et de conformité spécifiés dans l'Engagement 4 avant de mettre leur système d'IA générative sur le marché ou de le mettre en service, et tout au long de son cycle de vie.

Les Signataires s'engagent à s'efforcer d'atteindre le meilleur équilibre possible entre l'efficacité, l'interopérabilité, la robustesse et la fiabilité des solutions de marquage et de détection telles que décrites dans les Mesures 3.1 à 3.4 ci-dessous. Des contraintes opérationnelles peuvent imposer des limites au temps et aux

ressources de calcul, aux coûts de mise en œuvre, ainsi qu'à la capacité de passage à l'échelle pour des contenus de très grande ou de très petite taille.

Mesure 3.1 : Efficacité

Les Signataires mettent en œuvre des solutions techniques de marquage et de détection qui, conjointement, sont adaptées à leur finalité et capables de permettre aux personnes physiques de distinguer le contenu généré ou manipulé par leur système d'IA, contribuant ainsi à la confiance et à l'intégrité de l'écosystème de l'information.

Ces solutions sont considérées comme efficaces lorsque les personnes physiques peuvent accéder aux résultats de la détection et en comprendre la signification. Aucune mesure d'évaluation quantitative n'est liée à cette Mesure ; celle-ci requiert au contraire une évaluation de la fonctionnalité de détection fondée sur l'utilisateur. Les Signataires sont encouragés à publier des spécifications afin de guider les parties prenantes en aval sur la manière d'afficher l'information de détection produite par leur solution de détection. Les Signataires peuvent mener des études de panel pertinentes, dans la mesure où elles sont pertinentes pour leur cas d'usage applicatif, afin de vérifier dans quelle mesure les personnes physiques exposées aux résultats de la détection les comprennent.

Mesure 3.2 : Fiabilité

Les Signataires mettent en œuvre des solutions de marquage et de détection qui, conjointement, atteignent un niveau élevé de fiabilité dans différents contextes attendus et à travers les cas d'usage, dans la mesure où cela est techniquement réalisable et en cohérence avec l'état de l'art.

La fiabilité renvoie à la capacité des solutions de marquage et de détection à identifier et à distinguer avec exactitude l'origine du contenu généré ou manipulé par IA par rapport à d'autres contenus. L'évaluation de la fiabilité comporte deux composantes : (i) le degré d'exactitude de la solution de détection en ce qui concerne la détection du marquage dans des conditions nominales où le contenu ne subit aucune altération, et (ii) la manière dont l'exactitude des solutions de marquage et de détection varie en fonction de la longueur, de la taille, de la diversité et de la sémantique du contenu.

Les Signataires mesurent l'exactitude de la détection du contenu généré ou manipulé par IA au moyen de métriques pertinentes pour le type de marquages, telles que le taux d'erreur de la détection. De faibles taux d'erreur sont démontrés sur une variété d'échantillons, comprenant à la fois du contenu généré ou manipulé par leur système d'IA avec leur propre marquage et d'autres contenus (incluant éventuellement du contenu généré ou manipulé par d'autres systèmes d'IA). Les données utilisées pour l'évaluation de la fiabilité ne doivent, dans la mesure du possible, pas avoir été utilisées durant l'entraînement et le développement du système d'IA.

Les Signataires veillent à ce que ces mesures d'exactitude soient effectuées et rapportées avec du contenu destiné à être généré ou manipulé par le système d'IA, de longueur, de taille, de diversité et de sémantique variées, afin de démontrer que la performance des solutions de marquage et de détection mises en œuvre se généralise à travers des contextes et des cas d'usage divers.

Mesure 3.3 : Robustesse

Les Signataires mettent en œuvre des solutions de marquage et de détection qui, ensemble, maintiennent les niveaux de performance escomptés dans des conditions variables, couvrant à la fois les altérations courantes et les attaques adverses, dans la mesure où cela est techniquement réalisable et en cohérence avec l'état de l'art. Cette exigence ne s'applique pas aux systèmes d'IA qui sont exceptionnellement soumis uniquement à une seule couche de marquage par métadonnées telle que spécifiée dans la Mesure 1.1.

Premièrement, les Signataires veillent à ce que leurs solutions de marquage et de détection soient robustes face aux opérations de traitement typiques susceptibles d'introduire des degrés variables de distorsion en fonction des paramètres sous-jacents. Les opérations de traitement typiques comprennent :

- a. les modifications sur place, par exemple l'ajout / la suppression de bruit, le filtrage, la (re)compression, la capture d'écran / la diffusion d'écran, l'amélioration de la voix, la substitution lexicale, les homoglyphes, le changement de format de fichier, etc. ;
- b. les mécanismes de désynchronisation, par exemple la mise en miroir, le recadrage, la mise à l'échelle vers le haut / le bas, la rotation, le changement de rapport d'aspect, le décalage de hauteur tonale (pitch shifting), l'étirement temporel, l'insertion / la suppression de caractères, la paraphrase, les cycles de traduction, etc. ;
- c. la survie au trou analogique, par exemple l'impression-et-numérisation (et la reconnaissance optique de caractères), la lecture et l'enregistrement audio, la captation vidéo d'écran à la caméra, etc.

En outre, les Signataires veillent à ce que la détection soit robuste face aux modifications localisées du contenu, par exemple si des utilisateurs décident de flouter des visages pour des raisons de confidentialité avant la soumission du contenu à la solution de détection du Signataire.

Deuxièmement, les Signataires veillent à ce que la robustesse adverse de leurs solutions de marquage et de détection soit évaluée en termes de résilience face à des comportements malveillants tels que les attaques de copie, de suppression, de régénération et de modification des marquages. Les Signataires mesurent la robustesse face aux comportements malveillants par la capacité des solutions de marquage et de détection à vérifier l'intégrité du marquage, c'est-à-dire à déterminer si une marque a été altérée ou modifiée afin de présenter de manière trompeuse l'origine du contenu. Les Signataires choisissent les comportements malveillants pris en compte comme des menaces plausibles dans le monde réel, en fonction du type de contenu, du type de marque et du contexte de déploiement et de diffusion du contenu. Les Signataires sont encouragés à appliquer des pratiques de cybersécurité standard, telles que les limites de débit, afin de prévenir et de contrer l'usage malveillant et les attaques contre leurs solutions de marquage et de détection. Les Signataires sont encouragés à mettre fréquemment à jour leur évaluation des menaces afin de rester à jour face aux évolutions du paysage des menaces.

Pour évaluer la robustesse face aux opérations de traitement typiques et aux comportements malveillants, les Signataires utilisent les mêmes métriques de performance que pour l'exigence de fiabilité de la Mesure 3.2.

Mesure 3.4 : Interopérabilité

Les Signataires mettent en œuvre des solutions de marquage et de détection qui fonctionnent de manière transparente à travers de multiples systèmes, acteurs, contextes et implémentations techniques afin de permettre la détection du contenu généré ou manipulé par IA, quelle que soit la technique de marquage déployée dans les différents systèmes d'IA, dans la mesure où cela est techniquement réalisable.

À la date de publication du présent Code, les normes et/ou meilleures pratiques pertinentes en matière d'interopérabilité restent à élaborer, à l'exception des métadonnées signées numériquement. Par conséquent, une mise en œuvre échelonnée des exigences d'interopérabilité est adoptée comme suit.

Lors de la phase initiale, à compter de l'entrée en application de l'article 50, paragraphe 2, du règlement sur l'IA (AI Act), les Signataires :

- a. adoptent les normes établies ou les meilleures pratiques pertinentes pour mettre en œuvre le marquage et la détection par métadonnées, permettant ainsi une détection ouverte de ces marquages ;

- b. fournissent à la Commission et aux parties prenantes pertinentes des informations publiques sur la manière d'intégrer et d'accéder aux solutions de détection pour le ou les autres types de techniques de marquage qu'ils ont mis en œuvre dans leur système d'IA. Afin de faciliter la diffusion de ces informations, les Signataires sont encouragés à fournir les informations pertinentes sur leurs sites web et à alimenter les registres accessibles au public qui pourraient être créés pour rationaliser les intégrations.
- c. œuvrent vers un niveau minimal d'interopérabilité entre les solutions de détection pour le filigranage qui réduise la charge et les coûts pour les Signataires, tout en assurant un accès plus transparent pour les utilisateurs finaux. Plus précisément, les Signataires mettent en œuvre une solution d'interopérabilité pour leurs mécanismes de détection d'ici au 2 février 2027 en mettant en œuvre un ou plusieurs des éléments suivants :
 - d. une méthode d'accès reposant sur une norme industrielle interopérable accessible au public, qui permet d'acheminer les requêtes de détection vers les mécanismes de détection pertinents, et d'interpréter les résultats de détection au moyen de solutions de détection partagées ou indépendantes du fournisseur interagissant avec des points de terminaison mettant en œuvre une telle API, y compris des solutions de détection qui prennent en charge des balises publiques (public signposts) sans nécessairement les exiger. Les Signataires qui mettent en œuvre cette approche ne doivent pas créer de trafic inutile vers d'autres services lorsque ce trafic peut être évité en lisant directement une solution d'interopérabilité connue (telle qu'une balise) ;
- ii) une balise publiquement lisible (signpost) ou un autre mécanisme interopérable dans le contenu généré ou manipulé par IA qui signalera au public quelle solution de détection utiliser sans avoir à exécuter toutes les solutions de détection des fournisseurs de systèmes d'IA. Les Signataires prennent des mesures raisonnables pour veiller à ce que le ou les mécanismes de détection associés à leur(s) technique(s) de marquage par balise ou à une autre solution d'interopérabilité soient mis à la disposition de la Commission et des parties prenantes pertinentes afin qu'ils puissent être intégrés dans leurs flux de travail de détection ;
- iii) une solution de détection de fournisseur partagée par un consortium de Signataires, qui soit accessible et ouverte à tout autre Signataire souhaitant la rejoindre, y compris les PME et les PMC, sous réserve de garanties appropriées pour la protection des informations techniques confidentielles. Toute solution de détection agrégée de ce type doit être indépendante des Signataires participant au consortium, capable de détecter les sorties générées ou manipulées par leurs systèmes d'IA générative, et apte à fonctionner de manière interopérable avec d'autres solutions de détection ;
- iv) une autre solution d'interopérabilité qui réalise entre les Signataires une interopérabilité comparable aux solutions ci-dessus.

La solution d'interopérabilité est mise en œuvre d'une manière qui ne porte pas atteinte à la robustesse des autres filigranes numériques intégrés et qui inclut des garanties contre la création de risques d'usage abusif.

- a. Afin de progresser vers une interopérabilité complète, sous réserve de leurs ressources et de leurs capacités, les Signataires coopèrent de bonne foi avec la Commission et les parties prenantes pertinentes en vue du développement de solutions d'interopérabilité et de normes d'interopérabilité améliorées et conformes à l'état de l'art. En particulier, les Signataires sont encouragés à participer au groupe de travail décrit dans la Mesure 3.5. Les Signataires sont également encouragés à rejoindre et/ou à soutenir les organismes de normalisation internationaux et européens, ainsi que les initiatives de consortium axées sur le développement de normes ouvertes de marquage et de détection des contenus qui contribuent à opérationnaliser les mesures envisagées dans la présente Section du Code.

Lors de ces phases ultérieures, les Signataires s'engagent à mettre en œuvre ces normes en temps utile à mesure qu'elles émergent, dans la mesure où cela est techniquement réalisable et nécessaire pour le cas

d'usage applicatif de leurs systèmes d'IA, afin de continuer à se conformer à l'exigence d'interopérabilité au titre de l'article 50, paragraphe 2, du règlement sur l'IA (AI Act).

Mesure 3.5 : Faire progresser l'état de l'art (facultatif)

Sous réserve de leurs capacités et de leurs ressources, les Signataires sont encouragés à investir dans la recherche et le développement scientifiques et à collaborer avec les autorités compétentes, les chercheurs, les organisations de la société civile et d'autres parties prenantes pertinentes afin de faire progresser l'état de l'art en matière de techniques de marquage et de mécanismes de détection des contenus générés et manipulés par IA.

De telles avancées de l'état de l'art comprennent, entre autres :

- a. faire progresser les techniques de marquage et les mécanismes de détection existants en ce qui concerne le niveau atteignable d'efficacité, de fiabilité, de robustesse et d'interopérabilité ;
- b. faire progresser les solutions d'interopérabilité existantes vers un plus haut degré d'interopérabilité ;
- c. contribuer à et promouvoir la normalisation des interfaces des solutions de marquage et de détection, ainsi que des techniques elles-mêmes lorsque cela est possible ;
- d. développer de nouvelles solutions de marquage ou de détection qui soient indépendantes du fournisseur d'IA, par exemple des mécanismes de détection opérationnels au regard des exigences de qualité ci-dessus à travers de multiples systèmes d'IA, y compris des solutions de détection accessibles au public pouvant être exécutées localement, sans dépendance à des services de détection tiers ;
- e. établir des meilleures pratiques industrielles et/ou des bancs d'essai publics (benchmarks) pour l'évaluation des solutions de marquage et de détection et des techniques sous-jacentes, y compris en contribuant à la création de jeux de données de référence ;
- f. organiser et/ou participer à des exercices de red-teaming afin d'évaluer les limites des solutions de marquage et de détection à la pointe de l'état de l'art, ainsi que des techniques sous-jacentes, et d'identifier de nouvelles déficiences.

Enfin, les Signataires sont encouragés à participer activement au groupe de travail qui sera mis en place dans le cadre du Code afin de faciliter les activités mentionnées ci-dessus, le partage de bonnes pratiques, la cohérence dans la mise en œuvre des Mesures du Code, et la collaboration entre les Signataires, les autres acteurs de la chaîne de valeur et les parties prenantes pertinentes au moyen de réunions régulières.

Engagement 4 : Tests, vérification et conformité

TEXTE JURIDIQUE : Article 50, paragraphes 2 et 5, et considérant 133 du règlement sur l'IA (AI Act)

Afin de remplir efficacement leurs obligations au titre de l'article 50, paragraphes 2 et 5, du règlement sur l'IA (AI Act), ainsi que les Engagements et Mesures spécifiés dans la présente Section du Code, et d'en démontrer le respect, les Signataires s'engagent à mettre en place, à tenir à jour et à appliquer des processus de conformité, de test, de vérification et de surveillance, tels que spécifiés dans les mesures suivantes.

Mesure 4.1 : Processus de conformité

Les Signataires documentent, mettent en œuvre et tiennent à jour, conformément à l'état de l'art, un processus de conformité décrivant à un niveau général la manière dont ils ont mis en œuvre les différentes Mesures de la présente Section du Code afin d'assurer le respect de l'article 50, paragraphes 2 et 5, du règlement sur l'IA (AI Act). Cette mesure est mise en œuvre de manière proportionnée, en tenant compte de la taille et des ressources du Signataire, en particulier en ce qui concerne les fournisseurs qui sont des PME et des PMC, afin de minimiser la charge et de simplifier la conformité.

Les Signataires peuvent démontrer leur conformité au moyen de processus et de documentation de conformité existants, dans la mesure où ceux-ci satisfont aux mesures de la présente Section du Code. Lorsque les Signataires s'appuient sur des solutions de marquage et de détection fournies par des tiers ou mises en oeuvre au niveau du modèle d'IA générative, les Signataires peuvent tirer parti de la documentation fournie par ces tiers, sans préjudice de la responsabilité ultime du Signataire d'assurer le respect de l'article 50, paragraphes 2 et 5, du règlement sur l'IA (AI Act). Les Signataires qui s'appuient sur des solutions et des normes de marquage et de détection en source ouverte documentent le recours à ces normes et outils, ainsi que les informations supplémentaires, uniquement en ce qui concerne les mesures de la présente Section du Code pour lesquelles les informations publiques font défaut.

Mesure 4.2 : Tests, vérification et surveillance

Avant de mettre leur système d'IA générative sur le marché ou de le mettre en service, et régulièrement par la suite, les Signataires testent la conformité de leurs solutions de marquage et de détection aux exigences et aux mesures spécifiées dans la présente Section du Code. Cela est réalisé dans des conditions représentatives d'une utilisation en conditions réelles, dans l'objectif d'assurer l'alignement sur l'état de l'art en évolution.

Les Signataires qui sont des fournisseurs en aval de systèmes d'IA générative peuvent s'appuyer sur les résultats des tests effectués par un fournisseur de modèle en amont ou par un fournisseur tiers de solutions de marquage et de détection, pour autant que ces solutions satisfassent aux exigences spécifiées dans la présente mesure et sans préjudice de la responsabilité ultime du Signataire d'assurer le respect de l'article 50, paragraphes 2 et 5, du règlement sur l'IA (AI Act).

Jusqu'à ce qu'émergent des méthodes d'évaluation des performances et des référentiels reconnus pour les techniques de marquage et les mécanismes de détection, en particulier les méthodes et référentiels approuvés par le Bureau de l'IA après consultation du Comité IA et de la taskforce décrite à la Mesure 3.5, les Signataires testent et rendent compte des performances de leur solution selon des référentiels internes et les meilleures pratiques du secteur. Les Signataires peuvent associer des experts indépendants aux tests de leurs solutions, notamment pour évaluer la robustesse face aux attaques adverses au moyen d'exercices de red teaming, et/ou réaliser ces tests et évaluations sous supervision réglementaire dans le cadre de bacs à sable réglementaires de l'IA, comme prévu à l'article 57 du règlement sur l'IA (AI Act).

Les Signataires surveillent toute lacune de conformité étayée identifiée en interne ou signalée par des tiers concernés et mettent en oeuvre les actions correctives de suivi appropriées.

Mesure 4.3 : Formation

Les Signataires déploient des efforts proportionnés pour dispenser une formation appropriée à leur personnel exerçant des fonctions pertinentes pour assurer le respect de l'article 50, paragraphes 2 et 5, et de l'article 4 du règlement sur l'IA (AI Act). Cela comprend le personnel ou les sous-traitants externes qui participent à la conception et au développement des systèmes d'IA des Signataires, et qui sont chargés de veiller à ce que les Engagements et Mesures spécifiés dans la présente Section du Code soient mis en oeuvre efficacement. Cette mesure est mise en oeuvre de manière proportionnée, en tenant compte de la taille et des ressources du Signataire, en particulier en ce qui concerne les fournisseurs qui sont des PME et des PMC.

Mesure 4.4 : Coopération avec les autorités de surveillance du marché

Conformément à l'article 74 du règlement sur l'IA (AI Act) et à l'article 7 du règlement (UE) 2019/1020 (le règlement sur la surveillance du marché), les Signataires coopèrent avec les autorités de surveillance du marché compétentes au titre du règlement sur l'IA (AI Act) afin de démontrer le respect de l'article 50, paragraphes 2 et 5, du règlement sur l'IA (AI Act) et de leurs Engagements au titre de la présente Section du Code. À leur demande motivée, lorsque cela est nécessaire à l'exercice des tâches pertinentes des autorités de surveillance du marché, les Signataires fournissent la documentation spécifiée à la Mesure 4.1 ainsi que l'accès à leurs solutions de marquage et de détection. L'article 78 du règlement sur l'IA (AI Act) s'applique aux informations obtenues dans le cadre des activités de surveillance du marché, garantissant que les secrets d'affaires et les informations confidentielles sont préservés conformément au règlement sur l'IA (AI Act) et au droit de l'Union applicable.

Les Signataires sont encouragés à coopérer avec le Bureau de l'IA et le Comité IA en fournissant des informations sur les technologies susceptibles d'être utilisées pour assurer la transparence tout au long de la chaîne de valeur et qui ont été évaluées comme conformes au Code par le Bureau de l'IA et le Comité IA, par exemple au moyen d'un répertoire de l'UE des normes et technologies reconnues disponibles.

Glossaire

Partout où la présente Section renvoie à un terme défini à l'article 3 du règlement sur l'IA (AI Act), la définition du règlement sur l'IA (AI Act) s'applique.

Les termes suivants, avec le sens indiqué, sont utilisés dans la présente Section du Code. Sauf indication contraire, toutes les variations grammaticales des termes définis dans le présent Glossaire sont réputées couvertes par la définition pertinente.

Trou analogique (analog hole)

Désigne une catégorie d'opérations de traitement de contenu dans lesquelles le contenu numérique est restitué sous une forme analogique, généralement pour être présenté à des personnes physiques, avant d'être re-numérisé. Un exemple simple est l'enregistrement, à l'aide d'une caméra numérique, d'un contenu vidéo numérique affiché sur un écran.

API

Interface de programmation d'application (Application Programming Interface) : une interface utilisable par machine vers un système d'IA ou un autre service logiciel.

Texte conteneurisé (containerised text)

Texte intégré dans un format de fichier structuré ou un conteneur, où le contenu est organisé selon des règles spécifiques définies par le format. Le conteneur peut inclure des métadonnées, des informations de mise en page et des détails d'encodage qui influencent la manière dont le texte est stocké, affiché et traité. Les exemples comprennent le texte contenu dans des fichiers PDF, des documents Word ou des fichiers HTML.

Mécanisme de désynchronisation (desynchronization mechanism)

Désigne une catégorie d'opérations de traitement de contenu dans lesquelles la position des échantillons numériques est modifiée (et possiblement aussi leur valeur). Un exemple simple est l'application d'un effet miroir à une image numérique. Cette définition doit être considérée en parallèle avec celle de la « modification sur place » (in-place modification).

Mécanisme de détection pour une technique de marquage (detection mechanism for a marking technique)

Mécanisme de détection des filigranes numériques ou de vérification des marquages de métadonnées signées numériquement qui ont été ajoutés intentionnellement par un fournisseur d'un système d'IA ou par un tiers (par exemple, un fournisseur de modèle). Différents fournisseurs d'un système d'IA peuvent partager le même mécanisme de détection, généralement lorsqu'ils utilisent une technique de marquage normalisée ou développent conjointement une solution de détection partagée. Toutefois, en général, un mécanisme de détection est spécifique à une technique de marquage donnée, c'est-à-dire que différentes techniques de filigranage auront différents mécanismes de détection.

Signature numérique (digital signature)

Signature cryptographique permettant la vérification sécurisée de l'authenticité du fournisseur et de l'intégrité du contenu signé.

Utilisateur final (end-user)

Personne physique exposée au contenu généré ou manipulé par le système d'IA.

Taux d'erreur (error rate)

Se rapporte à la probabilité que le mécanisme de détection classe à tort certains contenus comme « générés ou manipulés par IA ». Les taux d'erreur peuvent être davantage décomposés pour une meilleure interprétabilité ; par exemple, le taux de faux positifs peut nuire à la confiance que l'on peut placer dans une technique de marquage et le mécanisme de détection associé.

Empreinte numérique (« fingerprinting »)

Également appelée « hachage perceptuel » (perceptual hashing) ou « hachage robuste » (robust hash). Technique d'indexation de contenu qui réduit le contenu à des descripteurs condensés pouvant être recherchés efficacement même si le contenu a fait l'objet d'altérations. L'empreinte numérique (« fingerprinting ») a été utilisée par le passé pour la reconnaissance ou l'identification de contenu fondées sur le contenu.

Détection forensique (forensic detection)

Détection de contenu généré ou manipulé par IA qui ne nécessite pas de marquage IA préalable. Par exemple, un mécanisme de détection forensique peut viser à capter une signature de signal intrinsèque présente dans le contenu généré ou manipulé par IA qui le rend différent d'un autre contenu.

Texte libre (free-form text)

Séquence brute de caractères sans structure, schéma ou format de conteneur englobant. Il n'est lié à aucune métadonnée, règle de mise en page ou encodage au-delà de la représentation de base des caractères. Des exemples pourraient être le texte affiché sur un site web ou au sein d'une conversation (chat).

Homoglyphes (homoglyphs)

Caractères dont la forme apparaît (presque) identique mais qui proviennent d'alphabets, de polices ou d'encodages différents. À ce titre, ils peuvent être exploités pour tromper les techniques de marquage de texte et les mécanismes de détection associés.

Modification sur place (in-place modification)

Désigne une catégorie d'opérations de traitement de contenu dans lesquelles la valeur des échantillons numériques, par exemple les pixels d'une image, est modifiée, mais non leur emplacement. Un exemple simple est l'augmentation du contraste d'une image numérique. Cette définition doit être considérée en parallèle avec celle de « mécanisme de désynchronisation » (desynchronization mechanism).

Journalisation (logging)

Enregistrement et indexation à l'identique (verbatim) du contenu (généralement du texte). Peut être utilisée pour une recherche rapide de contenu connu, c'est-à-dire qu'un référentiel d'entrées journalisées peut être interrogé pour déterminer si un contenu est connu pour avoir été généré ou manipulé par IA.

Marquage (marking)

Ajout ou intégration d'un marqueur à un contenu généré ou manipulé par IA, tel que des métadonnées signées numériquement ou un filigrane numérique imperceptible. L'objectif de ce marqueur est de faciliter la détection du contenu généré ou manipulé par IA.

Information de provenance (provenance information)

Enregistrement numérique relatif à un élément de contenu généré ou manipulé par un système d'IA qui en indique l'origine, la manière et le moment où le contenu a été généré ou manipulé, et le traitement appliqué au contenu.

Capture d'écran vidéo (screencasting)

Processus consistant à enregistrer numériquement la sortie visuelle affichée sur un écran (ordinateur, tablette, smartphone), souvent accompagnée d'une narration audio, et à la sauvegarder dans un fichier ou à la diffuser en direct.

Solution de balisage (signpost solution)

Technique de marquage ouvertement divulguée au sein du contenu et mécanisme de détection associé, utilisée pour indiquer quel mécanisme de détection (de filigrane numérique) doit être employé. Il s'agit d'un mécanisme d'interopérabilité destiné à rationaliser le flux de travail de détection et à éviter de soumettre chaque contenu faisant l'objet d'une vérification à l'ensemble des mécanismes de détection employés par les fournisseurs de systèmes d'IA générative lorsque les métadonnées ne sont pas disponibles. Elle peut être mise en œuvre, par exemple, sous la forme d'un marquage imperceptible lisible par machine.

Utilisateur (user)

Soit un déployeur au sens de l'article 3, point 4), du règlement sur l'IA (AI Act), soit une autre personne qui utilise le système d'IA générative d'un fournisseur.

UX

« Expérience utilisateur » (User Experience), c'est-à-dire la manière dont un utilisateur interagit avec les aspects d'un produit logiciel côté utilisateur et les perçoit.

Texte très court (very short text)

Texte si court que, dans de nombreux cas, il ne peut pas être marqué par filigrane numérique, même avec un niveau de fiabilité élémentaire. Au moment de la publication du Code, les techniques de l'état de l'art permettent le filigranage avec au moins un niveau de fiabilité élémentaire pour un texte aussi court que 200 jetons (tokens), avec l'attente que ce seuil diminuera à mesure que de nouvelles méthodes deviendront disponibles. D'ici là, l'expression « texte très court » doit être comprise comme un texte de moins de 200 jetons (tokens).

Filigrane numérique (watermark)

Technique qui intègre un marqueur au sein du contenu d'une manière qui est imperceptible (pour les personnes physiques) et robuste (le filigrane numérique est récupérable après altération du contenu). Le filigrane numérique est intrinsèquement lié à la représentation numérique du contenu (par exemple, les pixels d'une image, les échantillons audio), c'est-à-dire qu'il ne constitue pas des métadonnées auxiliaires. Le

filigranage a été utilisé par le passé pour la lutte contre le piratage, la mesure d'audience et la preuve de propriété.

Section 2 : Étiquetage des hypertrucages (« deepfakes ») et des textes publiés générés et manipulés par IA applicable aux déployeurs de systèmes d'IA (Article 50(4) et (5) du règlement sur l'IA (AI Act))

Objectifs

L'objectif global du présent Code de bonnes pratiques (« Code ») est d'améliorer le fonctionnement du marché intérieur, de promouvoir l'adoption d'une intelligence artificielle (« IA ») centrée sur l'humain et digne de confiance, tout en garantissant un niveau élevé de protection de la santé, de la sécurité et des droits fondamentaux consacrés par la Charte, y compris la démocratie, l'État de droit et la protection de l'environnement, contre les effets nuisibles de l'IA dans l'Union, et de soutenir l'innovation conformément à l'article 1(1) du règlement sur l'IA (AI Act).

Pour atteindre cet objectif global, les objectifs spécifiques de la présente Section du Code sont :

- a. servir de document d'orientation pour démontrer le respect des obligations des déployeurs de systèmes d'IA générative prévues à l'article 50(4) et (5) du règlement sur l'IA (AI Act), tout en reconnaissant que l'adhésion au Code ne constitue pas une preuve concluante du respect de ces obligations au titre du règlement sur l'IA (AI Act) ;
- b. garantir que les déployeurs de systèmes d'IA qui génèrent ou manipulent un contenu image, audio ou vidéo constituant un hypertrucage (« deepfake »), ou un texte publié dans le but d'informer le public sur des questions d'intérêt public, respectent leurs obligations au titre de l'article 50(4) et (5) du règlement sur l'IA (AI Act), et
- c. permettre aux autorités de surveillance du marché compétentes d'évaluer le respect, par les déployeurs qui s'appuient sur le Code pour démontrer la conformité avec ces obligations au titre de l'article 50(4) et (5) du règlement sur l'IA (AI Act), de manière cohérente, prévisible et uniforme dans l'ensemble de l'Union.

Considérants

Considérant que :

- a. **Divulgaration claire et distinguable** : Les Signataires reconnaissent que les progrès de l'IA générative permettent aux déployeurs de systèmes d'IA de générer ou de manipuler des contenus qui paraissent de plus en plus réalistes, rendant plus difficile pour les personnes de les distinguer d'un contenu authentique. Cela met en évidence la nécessité de la transparence pour protéger la confiance du public, préserver le discours démocratique et éclairer la prise de décision individuelle. Avec l'aide de systèmes d'IA, les déployeurs peuvent générer ou manipuler des hypertrucages (« deepfakes ») qui ressemblent à des personnes, objets, lieux, entités ou événements existants, paraissant faussement authentiques ou véridiques aux personnes physiques, ainsi que des textes sur des questions d'intérêt public publiés sans relecture humaine ni contrôle éditorial. La génération et la diffusion de tels contenus posent des risques spécifiques de tromperie et de manipulation, sapant la confiance dans l'écosystème de l'information et menaçant son intégrité. Les Signataires reconnaissent que la divulgation claire et distinguable de l'origine artificielle ou de la modification de tels contenus, requise au titre de l'article 50(4) et (5) du règlement sur l'IA (AI Act), constitue une garantie nécessaire pour sensibiliser les personnes et éclairer leur prise de décision. Cette garantie ne vise pas à fournir aux personnes des informations sur la fiabilité du contenu, mais plutôt à les informer de son origine artificielle ou du rôle du système d'IA dans la manipulation du contenu.

- b. **Mise en œuvre cohérente et efficace** : Pour rendre cette divulgation claire et distinguable pour toutes les personnes physiques, les Signataires reconnaissent la nécessité d'un certain niveau d'uniformité dans les méthodes de divulgation telles qu'énoncées dans les Engagements de la présente Section du Code. De telles méthodes doivent être efficaces et alignées sur l'état de l'art technique (y compris pour l'expérience utilisateur) dans la mesure où cela est faisable et approprié au contexte de diffusion. En particulier, une icône UE optionnelle offre une voie simple, pratique et uniforme, au titre du Code, pour le respect de l'article 50(4) et (5) du règlement sur l'IA (AI Act). Dans le même temps, les Signataires reconnaissent le large champ d'application transversal de l'article 50(4) du règlement sur l'IA (AI Act), couvrant une multitude de cas d'usage et de contextes à travers les secteurs, médias, plateformes et modalités. Les Signataires reconnaissent la nécessité d'un niveau minimal d'uniformité et de cohérence concernant la mise en œuvre des spécifications de conception et de placement de l'icône ou des labels équivalents. Dans le même temps, les Signataires peuvent préciser davantage la manière de mettre en œuvre les mesures envisagées dans la présente Section du Code dans des bonnes pratiques sectorielles conformes à l'article 50 du règlement sur l'IA (AI Act) et à la présente Section du Code.
- c. **Accessibilité** : Les Signataires reconnaissent l'importance de garantir une divulgation accessible des hypertrucages (« deepfakes ») et des textes publiés sur des questions d'intérêt public générés ou manipulés par IA relevant du champ d'application de l'article 50(4) du règlement sur l'IA (AI Act), en tenant compte des besoins des différents groupes de personnes vulnérables. La divulgation devrait être conçue pour garantir qu'elle soit aisément perceptible et compréhensible par toutes les personnes physiques, y compris les personnes en situation de handicap, les enfants et les personnes âgées.
- d. **Sensibilisation et responsabilité** : Les Signataires reconnaissent que les processus internes, les mesures de littératie et de sensibilisation, ainsi que les mécanismes de révision sont essentiels pour garantir une divulgation efficace des hypertrucages (« deepfakes ») et des textes publiés sur des questions d'intérêt public générés ou manipulés par IA relevant du champ d'application de l'article 50(4) du règlement sur l'IA (AI Act). Pour y parvenir, les Signataires s'engagent à étiqueter de tels contenus générés et manipulés par IA et à les divulguer au plus tard lors de la première exposition, sur la base des spécifications de conception et de placement figurant dans le présent Code. Compte tenu des ressources organisationnelles et des capacités techniques dont disposent les Signataires, cela inclut également des processus et mesures internes proportionnés visant à garantir que les personnes soient informées de la nature artificielle du contenu auquel elles sont exposées, y compris en rendant publiquement disponibles des informations sur l'utilisation par le Signataire de l'icône UE ou de labels équivalents et sur leur signification.
- e. **Promotion d'un écosystème de transparence et d'innovation** : Les Signataires reconnaissent que le respect de l'article 50(4) du règlement sur l'IA (AI Act) n'est pas un exercice de cases à cocher, mais vise à créer un écosystème de transparence qui s'articule avec les pratiques organisationnelles, les développements technologiques et la sensibilisation des personnes. Il poursuit l'objectif commun de donner aux personnes les moyens de prendre des décisions éclairées en leur permettant de comprendre l'origine et la nature du contenu de synthèse. Dans un tel écosystème, les Signataires reconnaissent que le Code permet des développements continus dans un paysage technologique en évolution rapide et à la lumière du champ d'application transversal de l'article 50(4) du règlement sur l'IA (AI Act). Les facteurs clés d'une telle structure du Code sont de construire un écosystème qui préserve l'uniformité et la reconnaissabilité, tout en permettant que de nouvelles données soient recueillies et adaptées en bonnes pratiques qui soutiennent le Code et le respect de l'article 50(4) et (5) du règlement sur l'IA (AI Act).

f. Garanties supplémentaires au titre d'autres dispositions du droit de l'Union et du droit

national : Les Signataires reconnaissent que les obligations de transparence au titre du règlement sur l'IA (AI Act) s'appliquent en parallèle de, et n'affectent pas, d'autres obligations ou exigences légales pouvant s'appliquer à la création, à la distribution ou à l'utilisation d'hypertrucages (« deepfakes ») et de textes publiés sur des questions d'intérêt public générés ou manipulés par IA au titre de la législation applicable de l'Union (par exemple, en matière de protection des données, de protection des consommateurs, de services numériques, de propriété intellectuelle, de droit des médias, de publicité politique, de droit pénal et d'autres cadres réglementaires pertinents).

Engagements

La présente Section du Code s'applique uniquement aux Signataires dans la mesure où ils sont déployeurs de systèmes d'IA qui génèrent ou manipulent un contenu image, audio ou vidéo constituant un hypertrucage (« deepfake »)¹ ou un texte publié dans le but d'informer le public sur des questions d'intérêt public relevant du champ d'application de l'article 50(4) du règlement sur l'IA (AI Act). Chaque référence ci-dessous à « contenu », « hypertrucage (deepfake) » ou « texte publié » implique un contenu qui est i) généré ou manipulé par IA et qui se qualifie comme hypertrucage (« deepfake ») au titre de l'article 3(60) du règlement sur l'IA (AI Act) (désigné comme « hypertrucage (deepfake) »), ou ii) un texte généré ou manipulé par IA publié dans le but d'informer le public sur des questions d'intérêt public, sans relecture humaine ni contrôle éditorial et pour lequel aucune personne physique ou morale ne détient la responsabilité éditoriale de la publication du contenu (désigné comme « texte publié » ou « texte publié sur des questions d'intérêt public »).

Compte tenu de la diversité des déployeurs et des contextes dans lesquels les hypertrucages (« deepfakes ») et les textes publiés peuvent être créés et diffusés, les Signataires et les associations dont ils font partie peuvent établir conjointement des bonnes pratiques communes pour leur secteur ou leur industrie, lesquelles pourraient fournir davantage de détails concernant la mise en œuvre des Engagements et des Mesures spécifiés dans la présente Section du Code. De telles bonnes pratiques devraient être pleinement alignées sur les mesures énoncées ci-dessous, sans empêcher les Signataires d'adapter les pratiques de divulgation à leurs secteurs spécifiques afin de tenir compte des particularités sectorielles et des règles existantes, ni de fournir une transparence supplémentaire, le cas échéant. Pour garantir la transparence, la responsabilité et permettre le suivi par les autorités de surveillance du marché compétentes, les bonnes pratiques sectorielles établies conformément au présent paragraphe sont rendues publiquement disponibles en ligne.

Dans la présente Section, les termes « will » (présent de l'indicatif à valeur d'obligation), « encouragé » et « peut » doivent être interprétés de la manière suivante :

1. le présent de l'indicatif à valeur d'obligation (« will ») est utilisé pour les mesures obligatoires au titre du Code qui doivent être satisfaites pour être conforme à l'article 50(4) ou (5) du règlement sur l'IA (AI Act) et qui sont suivies par les autorités de surveillance du marché compétentes ;
2. « encouragé » est utilisé pour les mesures optionnelles au titre du Code qui ne sont pas juridiquement requises et sont purement volontaires mais néanmoins recommandées ;
3. « peut » est utilisé pour les mesures optionnelles au titre du Code qui ne sont pas juridiquement requises et sont purement volontaires ou qui offrent aux déployeurs une flexibilité quant à la manière de mettre en œuvre la mesure ou l'engagement concerné.

¹ Un hypertrucage (« deepfake ») est défini à l'article 3(60) du règlement sur l'IA (AI Act) comme un contenu image, audio ou vidéo généré ou manipulé par IA qui ressemble à des personnes, objets, lieux, entités ou événements existants et paraîtrait faussement authentique ou véridique à une personne.

Engagement 1 : Divulgence des hypertrucages (« deepfakes ») et des textes publiés

TEXTE JURIDIQUE : article 50, paragraphes 4 et 5, et considérants 133 et 135 du règlement sur l'IA (AI Act)

4. Les déployeurs d'un système d'IA qui génère ou manipule des images ou des contenus audio ou vidéo constituant un hypertrucage (« deepfake ») indiquent que les contenus ont été générés ou manipulés par une IA. Cette obligation ne s'applique pas lorsque l'utilisation est autorisée par la loi à des fins de détection, de prévention, d'enquête ou de poursuite d'une infraction pénale. Lorsque le contenu fait partie d'une œuvre ou d'un programme manifestement artistique, créatif, satirique, fictif ou analogue, les obligations de transparence énoncées au présent paragraphe se limitent à la divulgation de l'existence de tels contenus générés ou manipulés d'une manière appropriée qui n'entrave pas l'affichage ou la jouissance de l'œuvre.

Les déployeurs d'un système d'IA qui génère ou manipule un texte publié dans le but d'informer le public sur des questions d'intérêt public indiquent que le texte a été généré ou manipulé par une IA. Cette obligation ne s'applique pas lorsque l'utilisation est autorisée par la loi à des fins de détection, de prévention, d'enquête ou de poursuite d'infractions pénales, ou lorsque le contenu généré par IA a fait l'objet d'un processus de contrôle humain ou de contrôle éditorial et lorsqu'une personne physique ou morale assume la responsabilité éditoriale de la publication du contenu.

5. Les informations visées aux paragraphes 1 à 4 sont fournies aux personnes physiques concernées de manière claire et reconnaissable au plus tard au moment de la première interaction ou exposition. Les informations sont conformes aux exigences applicables en matière d'accessibilité.

Pour remplir leurs obligations au titre de l'article 50, paragraphes 4 et 5, du règlement sur l'IA (AI Act), les Signataires s'engagent à assurer une divulgation cohérente et efficace de l'origine artificielle des hypertrucages (« deepfakes ») ou des textes publiés.

Les Signataires s'engagent à mettre en œuvre cette divulgation au moyen de l'icône européenne disponible fournie à l'Annexe 1 du Code, ou au moyen d'une icône ou d'un label équivalent conforme aux spécifications de conception et de placement décrites dans les mesures suivantes.

Les Signataires peuvent intégrer l'icône ou le label équivalent dans leurs pratiques de divulgation existantes au titre de la législation sectorielle applicable de l'Union et nationale ou des codes de conduite professionnels, dans la mesure où ces pratiques sont conformes à la présente section du Code et à l'article 50, paragraphes 4 et 5, du règlement sur l'IA (AI Act).

Les Signataires reconnaissent que l'acte d'étiquetage ne les dispense pas des autres lois de l'Union et des États membres, telles que celles relatives à la protection des droits et libertés des tiers, y compris les exigences légales applicables concernant l'obtention du consentement des personnes représentées ou des titulaires de droits.

Mesure 1.1 : Spécifications de conception

Lorsqu'une divulgation visuelle est possible, les Signataires mettent en œuvre les spécifications de conception suivantes :

- a. L'icône ou le label équivalent comprend, comme élément visuel principal, l'acronyme en majuscules « AI » en langue anglaise (par exemple, une icône « AI »), à moins que l'usage de l'anglais ne soit incompatible avec les lois nationales applicables relatives à l'emploi des langues en matière commerciale ou administrative, auquel cas l'acronyme peut être divulgué dans la langue nationale. Les lettres de l'acronyme ont la même dimension verticale. En cas de redimensionnement, les proportions des lettres doivent être préservées.
- b. En outre, les Signataires sont encouragés à compléter l'acronyme figurant dans l'icône par des informations indiquant si l'hypertrucage (« deepfake ») ou le texte publié est manipulé ou généré par une IA dans une seconde couche (interactive) lorsque cela est techniquement réalisable, dans l'icône ou à côté de l'icône (par exemple, un texte indiquant « modifié » ou « généré », tel qu'illustré à l'Annexe 1). Les Signataires sont en outre encouragés à divulguer dans la seconde couche (interactive) ce qui a été modifié par le système d'IA (par exemple, un texte ou un pictogramme décrivant qu'un visage a été altéré).
- c. L'icône ou le label équivalent peut apparaître dans différentes tailles selon le contexte, dès lors que les Signataires veillent à ce que la divulgation soit claire et reconnaissable pour les personnes physiques. L'icône ou le label équivalent peut être exprimé dans différents styles (par exemple, rapport de contraste, couleur ou typographie), dès lors qu'il demeure clair, accessible et reconnaissable, c'est-à-dire lisible et identifiable par les personnes physiques.

Une icône européenne respectant les spécifications de conception pour la divulgation visuelle est mise à la disposition du public et peut être utilisée librement par tous (voir Annexe 1).

Lorsqu'une divulgation visuelle n'est pas possible (par exemple, un contenu audio uniquement), les Signataires mettent en œuvre les spécifications de conception suivantes :

- a. La divulgation comprend, au début de l'hypertrucage (« deepfake ») lui-même, une brève clause de non-responsabilité audible en langage naturel clair et simple, soit dans la même langue que le contenu, soit en anglais, divulguant de manière perceptible l'origine artificielle de l'hypertrucage audio.
- b. Le cas échéant, la clause de non-responsabilité audio est complétée par des informations indiquant si le contenu est généré ou manipulé par une IA. Les Signataires sont en outre encouragés à divulguer ce qui a été modifié par le système d'IA.

À titre alternatif, les Signataires peuvent développer et recourir à d'autres formes de divulgations audibles (par exemple, un earcon) en attendant le développement d'une solution audio commune à l'échelle de l'UE par le Groupe de travail (« Task Force »), décrit à la Mesure 1.3, si le déploiement d'une telle solution de divulgation audible s'accompagne de mesures de sensibilisation appropriées (par exemple, campagnes d'information du public, avis explicatifs ou clauses de non-responsabilité répétés) afin d'assurer une large compréhension de la solution par les personnes physiques exposées à l'hypertrucage audio.

Lorsqu'ils appliquent les spécifications de conception de la présente mesure, les Signataires tiennent compte de la composition potentiellement diverse du public exposé au contenu (y compris des niveaux divergents de littératie en matière d'IA et de littératie numérique, de maîtrise de la langue ou de connaissances générales, ainsi que des catégories d'utilisateurs vulnérables tels que les enfants et les personnes âgées), et de la nature potentiellement sensible du contexte dans lequel le contenu est utilisé et diffusé (par exemple, dans les secteurs financier, médical, éducatif ou d'autres secteurs sensibles).

Les Signataires assurent une divulgation accessible au regard des différentes modalités et contextes de contenu, conformément au droit de l'Union applicable, en particulier la directive (UE) 2019/882 (l'acte européen sur l'accessibilité) et la directive (UE) 2016/2102 (la directive sur l'accessibilité des sites web). Cela comprend, sans s'y limiter, l'application de :

- a. descriptions audio ou indices alternatifs pour les éléments de divulgation visuelle ;

- b. indices tactiles et haptiques pour les contenus audio uniquement, en tenant compte des besoins des utilisateurs finaux ayant des déficiences auditives (par exemple, une alerte par vibration avant la lecture de l'audio indiquant que l'audio contient un contenu d'hypertrucage) ;
- c. icônes ou labels équivalents à contraste élevé et compatibilité avec les lecteurs d'écran, y compris pour les utilisateurs finaux présentant des déficiences de la vision des couleurs ;
- d. détectabilité de l'icône ou du label équivalent par les technologies d'assistance.

Les Signataires sont encouragés à mettre en œuvre toute norme ou ligne directrice d'accessibilité pertinente disponible, y compris, sans s'y limiter, la norme harmonisée ETSI EN 301 549 « Exigences d'accessibilité pour les produits et services TIC » ou les Règles pour l'accessibilité des contenus web 2.1 du W3C, dans la mesure où les services ou produits des Signataires relèvent du champ d'application des exigences d'accessibilité applicables et de ces normes ou lignes directrices.

Les Signataires qui sont également fournisseurs de plateformes en ligne ou de moteurs de recherche en ligne, y compris les très grandes plateformes en ligne ou les très grands moteurs de recherche en ligne désignés et soumis à l'obligation de mettre en place des mesures d'atténuation des risques au titre de l'article 35, paragraphe 1, du règlement sur les services numériques (DSA), sont encouragés à utiliser l'icône européenne ou une autre solution conforme équivalente afin de permettre aux déployeurs de systèmes d'IA de mettre en œuvre leur obligation d'étiquetage d'une manière conforme à la présente section du Code et à l'article 50, paragraphe 4, du règlement sur l'IA (AI Act) (par exemple, en mettant à disposition des solutions de divulgation dans les interfaces de téléversement de la plateforme à l'intention des déployeurs).

Mesure 1.2 : Spécifications de placement

Pour satisfaire aux exigences légales d'étiquetage de manière claire et reconnaissable au plus tard au moment de la première exposition au titre de l'article 50, paragraphe 5, du règlement sur l'IA (AI Act), et compte tenu de la diffusion des textes publiés et des hypertrucages (« deepfakes ») en particulier sur les plateformes en ligne ou les moteurs de recherche, dans des environnements tant hors ligne qu'en ligne, les Signataires appliquent les spécifications de placement décrites dans la présente mesure.

Sous-mesure 1.2.1 : Principes généraux relatifs aux spécifications de placement

Les Signataires appliquent les principes généraux suivants, applicables à toutes les modalités et à tous les contextes de contenu, à l'exception des hypertrucages (« deepfakes ») qui font partie d'œuvres artistiques, créatives, satiriques, fictives ou analogues qui sont soumises au régime de divulgation spécifique précisé à l'Engagement 3 :

- a. Compte tenu du format du contenu et du contexte de diffusion, l'icône ou le label équivalent est placé de manière appropriée et perceptible garantissant une reconnaissance immédiate par les personnes physiques sans nécessiter d'(inter)action de l'utilisateur ni une attention soutenue.
- b. L'icône ou le label équivalent reste visible au moins pendant une durée suffisante pour être remarqué dans des conditions d'exposition normales, afin d'en assurer la perceptibilité.
- c. L'icône ou le label équivalent est directement intégré au contenu, à moins que des alternatives équivalentes à une icône intégrée ne soient disponibles (par exemple, une superposition d'interface utilisateur qui, pour les personnes physiques, semble se trouver sur le contenu), dès lors que les Signataires respectent les spécifications de placement de la présente mesure et s'efforcent, dans toute la mesure du possible, de garantir que la divulgation de l'hypertrucage (« deepfake ») ou du texte publié tient compte de la chaîne de distribution et de diffusion du contenu.

- d. L'icône ou le label équivalent est clairement perceptible et reconnaissable au plus tard au moment de la première exposition d'une personne physique à l'hypertrucage (« deepfake ») ou au texte publié (par exemple, en maintenant un espacement suffisant par rapport aux autres éléments de superposition, aux clauses de non-responsabilité et aux éléments sonores/visuels à l'écran/à l'affichage, et en demeurant visible sur tout arrière-plan).

Les Signataires sont encouragés à collaborer, dans la mesure du possible, avec les acteurs dont ils utilisent les services ou produits pour distribuer ou diffuser plus avant l'hypertrucage (« deepfake ») ou le texte publié (par exemple, éditeurs, plateformes en ligne ou commerce de détail) afin de préserver l'icône ou le label équivalent dans le but de garantir que la divulgation requise accompagne l'hypertrucage (« deepfake ») ou le texte publié tout au long de sa distribution ultérieure, qu'elle soit en ligne ou hors ligne, et sur l'ensemble de la chaîne de création, de distribution et de diffusion du contenu. Cet encouragement s'applique également aux divulgations relevant des Engagements 3 et 4 de la présente section du Code.

Sous-mesure 1.2.2 : Spécifications de placement lorsqu'une divulgation visuelle est possible

Outre les principes généraux, les Signataires mettent en œuvre les spécifications de placement suivantes lorsqu'une divulgation visuelle est possible :

- a. L'icône ou le label équivalent apparaît à un endroit approprié où n'existe aucun élément de superposition interférant (par exemple, dans le coin supérieur droit d'un hypertrucage image ou vidéo).
- b. Pour tenir compte des situations où différentes personnes physiques peuvent être exposées à un hypertrucage vidéo à des moments différents (par exemple, un contenu en direct) et pour tenir compte de l'utilisation en aval (par exemple, captures d'écran et fragments découpés), les Signataires affichent l'icône ou le label équivalent au début de la vidéo ainsi que, lorsque cela est possible, à intervalles réguliers tout au long de la vidéo et, au minimum, après les interruptions (par exemple, après les pauses commerciales ou publicitaires). Lorsque cela est possible et approprié selon le contexte, les Signataires sont en outre encouragés à afficher l'icône ou le label équivalent tout au long de l'hypertrucage vidéo ou pendant la partie hypertrucage de la vidéo, lorsque le segment d'hypertrucage ne couvre pas nécessairement l'intégralité de la durée de la vidéo.
- c. Lorsqu'un hypertrucage (« deepfake ») est exclusivement utilisé dans le cadre d'un contexte professionnel interne fermé (par exemple, à des fins de formation ou d'information des employés), la divulgation peut être placée dans l'interface utilisateur, le cadre physique ou tout autre support approprié facilement accessible aux personnes physiques exposées à l'hypertrucage, les informant avant qu'elles ne soient exposées à l'hypertrucage, dès lors qu'elle est claire et reconnaissable pour les personnes physiques impliquées dans le cadre professionnel.
- d. Pour les hypertrucages visuels, les divulgations audibles ne peuvent être mises en œuvre qu'à titre de méthode de divulgation supplémentaire et sont toujours accompagnées de divulgations visuelles (par exemple, lorsque l'audio fait partie de l'expérience utilisateur ou selon ce qui est approprié à des fins d'accessibilité).
- e. Pour les hypertrucages audio, lorsqu'un écran est disponible, une divulgation visuelle supplémentaire fondée sur l'icône ou le label équivalent est mise à disposition (par exemple, lorsqu'un écran est disponible dans une voiture ou sur l'affichage d'un smartphone), en complément de la clause de non-responsabilité audible.

f. Pour les textes publiés, les Signataires placent l'icône ou le label équivalent, par exemple au-dessus ou en haut du texte, près du titre du texte, ou dans le colophon au début du texte, dès lors que le placement est clair, cohérent et reconnaissable pour l'utilisateur final. Le cas échéant, les Signataires peuvent étiqueter uniquement la partie du texte qui est générée ou manipulée par IA. Pour les textes courts (mots isolés ou brèves phrases), lorsque l'étiquetage des sorties textuelles dégraderait la lisibilité et l'utilisabilité, les Signataires doivent néanmoins procéder à l'étiquetage mais peuvent assurer la divulgation au moyen d'un avis contextuel dans l'interface utilisateur (par exemple, un indicateur adjacent à la sortie, ou en divulguant le recours à une IA au début d'une session d'exposition de l'utilisateur ou lors de la première interaction avec le contenu).

Sous-mesure 1.2.3 : Spécifications de placement lorsqu'une divulgation visuelle n'est pas possible

Outre les principes généraux, lorsqu'une divulgation visuelle n'est pas possible, les Signataires incluent une clause de non-responsabilité audible au plus tard au moment de la première exposition à l'hypertrucage (« deepfake »). En outre, des indices audibles peuvent être utilisés (par exemple, des tonalités ou des earcons) pour faciliter la reconnaissance.

Afin de prendre en compte l'utilisation en aval (par exemple, des fragments audio découpés) ainsi que l'exposition de longue durée ou en direct, les Signataires complètent, le cas échéant, la clause de non-responsabilité audible initiale au début du contenu par des rappels à intervalles réguliers (par exemple, clauses de non-responsabilité, tonalités ou earcons) pendant toute la durée de l'hypertrucage audio afin d'assurer une sensibilisation continue, et au minimum après les interruptions (par exemple, par des publicités).

Mesure 1.3 : Participation volontaire à un groupe de travail au titre du Code

Les Signataires sont encouragés à soutenir les travaux et activités d'un groupe de travail (« Task Force ») dédié au titre du Code, visant à faire progresser le développement et l'utilisabilité de l'icône européenne en tant que mise en œuvre minimale conforme à l'état de l'art. Le groupe de travail vise, entre autres, à :

- a. promouvoir davantage l'adoption et le développement de l'icône européenne (y compris les versions non visuelles/audio uniquement de l'icône) ;
- b. mettre à jour l'icône européenne afin d'assurer en tout temps une divulgation conforme à l'état de l'art, y compris en cohérence avec les normes d'expérience utilisateur (UX) ;
- c. évaluer la faisabilité et œuvrer à faire progresser les solutions techniques, y compris le développement d'une seconde couche interactive liée à l'icône européenne ou à d'autres labels équivalents, qui offrent des possibilités d'intégration de l'icône européenne ou du label équivalent dans les services de diffusion de contenu en ligne afin de fournir aux utilisateurs finaux des informations supplémentaires (par exemple, le type d'implication de l'IA dans un hypertrucage (« deepfake ») ou un texte publié). Cette seconde couche interactive peut être mise en œuvre dans la mesure où cela est faisable et pratique, et en cohérence avec les solutions techniques de marquage décrites à la Section 1 du Code, afin d'assurer la robustesse des informations fournies dans l'icône et des informations qui l'accompagnent ;
- d. œuvrer à l'identification de méthodes communes pour divulguer le type d'implication de l'IA dans le cas des hypertrucages (« deepfakes ») ou des textes publiés manipulés par IA, en vue de faire progresser la cohérence dans la mise en œuvre entre les Signataires ;
- e. offrir un forum permettant aux Signataires d'échanger de bonnes pratiques entre secteurs et, éventuellement, de développer des bonnes pratiques sectorielles concernant la divulgation des hypertrucages (« deepfakes ») et des textes publiés, en cohérence avec la présente section du Code.

Les Signataires et les parties prenantes concernées de divers secteurs, y compris, sans s'y limiter, les chercheurs indépendants, les organisations de la société civile et les autorités nationales compétentes, sont encouragés à participer au groupe de travail.

Engagement 2 : Processus internes

Afin de remplir efficacement les obligations prévues à l'article 50, paragraphes 4 et 5, du règlement sur l'IA (AI Act), ainsi que les Engagements et Mesures spécifiés dans la présente Section du Code, et de démontrer leur respect, les Signataires s'engagent à mettre en place ou à maintenir des processus internes, des mesures de sensibilisation et des mécanismes de réexamen tels que spécifiés dans les mesures ci-dessous et proportionnés à leur taille et à leurs ressources disponibles.

Mesure 2.1 : Processus interne de conformité

Proportionnellement à leur taille, à leurs ressources organisationnelles disponibles et à leurs capacités, les Signataires mettent en place ou maintiennent des processus internes de conformité et une documentation appropriés qui précisent la manière dont ils mettent en oeuvre les obligations de divulgation à l'aide de l'icône ou d'un label équivalent. Cette documentation peut inclure une description générale ainsi que des exemples représentatifs, concrets et réels de la manière dont les divulgations sont mises en oeuvre en pratique, conformément aux Engagements 1, 3 et 4.

Les Signataires sont encouragés à publier des informations sur la solution de divulgation qu'ils appliquent d'une manière claire et accessible. Si les Signataires utilisent l'icône de l'UE accessible au public au titre de l'Engagement 1, ils peuvent le faire en utilisant les informations fournies avec l'icône de l'UE. Si un Signataire choisit d'appliquer sa propre icône ou un label équivalent, il est encouragé à rendre publiquement accessible une description similaire de l'icône ou du label. Cela peut inclure des images de l'icône ou du ou des labels équivalents utilisés, ainsi que des explications sur la signification de chaque label. Lorsqu'une icône ou un label équivalent revêt une signification différente selon le contexte, il est encouragé que toute variation soit expliquée.

Lorsque cela est approprié, et en particulier en cas d'utilisation régulière de systèmes d'IA pour créer des hypertrucages (« deepfakes ») ou des textes publiés, les Signataires mettent en place ou maintiennent un processus visant à garantir et à vérifier que les spécifications de conception et de placement sont correctement mises en oeuvre afin d'atténuer les risques de contenu non étiqueté ou incorrectement étiqueté.

Les fournisseurs de services de médias au sens de l'article 2, paragraphe 2, du règlement (UE) 2024/1083 peuvent se conformer à la présente Mesure en appliquant leurs procédures existantes et leurs normes professionnelles établies, dans la mesure où ces pratiques sont conformes aux mesures du présent Engagement et à l'article 50, paragraphes 4 et 5, du règlement sur l'IA (AI Act).

Mesure 2.2 : Sensibilisation et maîtrise

Les Signataires déploient des efforts proportionnés à leur taille, à leurs ressources organisationnelles disponibles et à leurs capacités afin d'assurer la sensibilisation aux obligations de divulgation prévues à l'article 50, paragraphes 4 et 5, du règlement sur l'IA (AI Act) au sein de leur personnel, y compris les employés, et des sous-traitants externes directement impliqués dans la mise en oeuvre des mesures de divulgation ou dans la surveillance du respect des mesures de la présente Section du Code.

Les Signataires sont encouragés à dispenser une formation ou des orientations équivalentes couvrant les situations dans lesquelles la divulgation est légalement requise, la manière dont les divulgations sont mises en oeuvre dans le flux de travail, les cas impliquant une responsabilité éditoriale, les cas impliquant une oeuvre artistique, créative, satirique, fictionnelle ou analogue, les considérations d'accessibilité, et les procédures de correction des labels manquants ou incorrects lorsque ceux-ci ont été identifiés.

Les efforts liés à la maîtrise de l'IA devraient être proportionnés à la taille et aux ressources disponibles du Signataire, et appliqués de manière appropriée aux rôles, aux connaissances techniques, à l'expérience et à la

formation du personnel impliqué dans la création, la modification et la diffusion des contenus pertinents soumis aux obligations d'étiquetage. Les Signataires demeurent libres de déterminer les formats de formation et leur fréquence.

Mesure 2.3 : Réexamen, retours d'information et coopération avec les autorités

Les Signataires soutiennent la mise en oeuvre efficace des spécifications de conception et de placement au moyen d'un réexamen interne et de retours d'information externes.

Les Signataires sont encouragés à fournir des canaux permettant aux personnes physiques ou aux tiers (par exemple, signaleurs de confiance, chercheurs indépendants, universitaires, vérificateurs de faits) de signaler les divulgations manquantes ou incorrectes, le cas échéant au moyen de mécanismes de signalement existants (par exemple, mécanismes de signaleurs de confiance, interfaces pour les services tiers de vérification des faits ou mécanismes de notification et d'action).

Les Signataires réexaminent les cas ayant fait l'objet d'un signalement étayé comme étant mal étiquetés ou incorrectement étiquetés et prennent des mesures pour remédier sans retard indu aux cas de non-respect de l'article 50, paragraphes 4 et 5. Les Signataires coopèrent avec les autorités compétentes conformément au droit de l'Union et au droit national applicables (par exemple, les autorités nationales de surveillance du marché).

Engagement 3 : Divulcation pour les oeuvres artistiques, créatives et similaires

Conformément à l'article 50, paragraphe 4, du règlement sur l'IA (AI Act), les Signataires s'engagent à mettre en oeuvre des mesures visant à divulguer les hypertrucages (« deepfakes ») qui s'inscrivent dans le cadre d'une oeuvre ou d'un programme manifestement artistique, créatif, satirique, fictionnel ou analogue, d'une manière qui n'entrave pas l'affichage ou la jouissance de l'oeuvre, y compris son exploitation et son utilisation normales, tout en préservant l'utilité et la qualité de l'oeuvre.

Dans ce contexte, les Signataires :

- a. utilisent une icône ou un label équivalent conformément aux spécifications de conception de la Mesure 1.1 et le placent d'une manière appropriée au type d'oeuvre artistique, créative, satirique, fictionnelle ou analogue et au contexte dans lequel elle est présentée.
- b. veillent à ce que cette divulgation et son placement soient clairs, distinguables et accessibles à toutes les personnes physiques, et fournis au plus tard au moment de la première exposition au contenu contenant l'hypertrucage (« deepfake ») (par exemple, dans les notes ou la description accompagnant fournies aux utilisateurs, dans le générique de début ou de fin, etc.).
- c. veillent à ce que l'icône ou le label équivalent soit perceptible pendant une durée suffisante pour être facilement remarqué dans des conditions normales de visionnage ou d'exposition, afin de garantir sa perceptibilité.

Lorsque le contenu d'hypertrucage (« deepfake ») est mis à disposition de manière numérique ou interactive (par exemple, sur des sites web, des applications ou d'autres interfaces utilisateur), l'icône ou le label équivalent peut être placé à l'extérieur mais à proximité du cadre de la vidéo ou de l'image, ou à proximité du contenu audio, et intégré dans des éléments de l'interface utilisateur ou des superpositions, sous le contrôle du Signataire. Les Signataires s'engagent à veiller à ce que toute solution de divulgation contextuelle de ce type soit perceptible par l'utilisateur final sans qu'il soit nécessaire d'effectuer des actions dédiées ou un engagement supplémentaire. Parmi les exemples de divulgation contextuelle figurent les divulgations faisant partie de l'interface utilisateur (par exemple, dans le cadre des sites web ou des interfaces d'application de lunettes connectées ou d'autres dispositifs numériques), de brèves notes, une icône ou un label non intrusif

qui, par un clic ou un survol, fournit davantage d'informations, ou des avertissements apparaissant à proximité du contenu.

Lorsque le contenu est mis à disposition de manière non numérique ou non interactive (par exemple, expositions, galeries d'art, cinémas, festivals ou contextes similaires, audio ou vidéo sur un support physique), les divulgations peuvent être fournies au point d'entrée ou de vente en ligne ou physique, dans le cadre des informations introductives ou d'accompagnement (par exemple, brochure d'exposition ou billet d'entrée), ou des informations fournies via un support physique (par exemple, l'emballage).

Engagement 4 : Réexamen humain et contrôle éditorial pour les textes publiés

Les Signataires qui sont des fournisseurs de services de médias au sens de l'article 2, paragraphe 2, du règlement (UE) 2024/1083 et qui sont déjà soumis à des normes éditoriales ainsi qu'à des cadres réglementaires, de corégulation ou d'autorégulation, tels qu'applicables à ces fournisseurs de services de médias en vertu du droit et des cadres de l'UE et nationaux en matière de médias, peuvent invoquer l'exception à l'obligation de divulgation prévue à l'article 50, paragraphe 4, deuxième alinéa, du règlement sur l'IA (AI Act) en appliquant leurs procédures de réexamen et éditoriales existantes et leurs normes professionnelles établies, selon le cas, pour se conformer au présent Engagement.

Tous les autres Signataires, y compris ceux ne disposant pas de telles procédures de réexamen ou éditoriales, s'engagent à établir, à adapter ou à maintenir des politiques appropriées de réexamen humain ou de contrôle éditorial préalablement à la publication, et à ce qu'une personne physique ou morale assume la responsabilité éditoriale de la publication. Ces politiques internes peuvent s'appuyer sur des processus existants, sont proportionnées à la taille et aux ressources du déployeur, et comprennent au moins les éléments suivants :

- a. L'identification de la personne physique ou morale assumant la responsabilité éditoriale (nom, fonction et coordonnées) ;
- b. Un aperçu des mesures organisationnelles concrètes ainsi que des ressources humaines, affectées pour garantir qu'un réexamen humain ou un contrôle éditorial adéquat est réalisé et que la responsabilité éditoriale est assumée avant la publication du texte publié. Cela n'implique pas de devoir documenter les cas individuels de réexamen humain ou de contrôle éditorial portant sur des publications de textes individuelles.

Lorsqu'elles ne sont pas déjà accessibles au public, les Signataires s'engagent à publier les coordonnées de la fonction, des personnes physiques ou des personnes morales assumant la responsabilité éditoriale, afin de garantir la responsabilité.

Les Signataires peuvent consigner des informations supplémentaires sur la nature du réexamen ou sur le type d'intervention du système d'IA dans le texte publié.

La mise en oeuvre du présent Engagement n'affecte en aucune manière la liberté des médias, l'indépendance éditoriale et la protection des informations relatives aux sources journalistiques.

Annexe

Annexe 1 : Icônes de l'UE accessibles au public

La présente Annexe contient les icônes de l'UE qui peuvent être utilisées par les déployeurs pour se conformer à la Mesure 1.1. Trois icônes principales sont fournies. Une icône est destinée à divulguer les hypertrucages (« deepfakes ») ou textes publiés entièrement générés par IA (Figure 1), une deuxième icône est destinée à divulguer les hypertrucages (« deepfakes ») ou textes publiés manipulés/partiellement modifiés par IA (Figure 2), et une troisième icône de base est incluse pour permettre aux déployeurs de la compléter par une couche interactive comportant davantage d'informations ou par un label textuel alternatif, le cas échéant (Figure 3).

La conception et le développement de l'icône de l'UE ont fait l'objet de tests empiriques auprès des utilisateurs menés dans plusieurs États membres, qui ont évalué la perception de l'authenticité, la visibilité et la reconnaissance, la compréhensibilité et la confiance. Les tests auprès des utilisateurs ont porté sur diverses conceptions d'icônes et ont démontré que, si toutes les conceptions testées atteignaient une visibilité et une fiabilité élevées, les variantes comportant un label textuel clair (à savoir « modifié ») obtenaient de bien meilleurs résultats en termes de visibilité et de clarté pour les utilisateurs finaux. Les répondants participant à l'enquête ont indiqué qu'un texte explicite sur la génération ou la manipulation par IA réduit l'ambiguïté quant à la nature du contenu. Par conséquent, sur la base de ces retours des utilisateurs, le Code inclut une icône pour le contenu entièrement généré par IA (« AI + GENERATED ») et une autre pour le contenu partiellement manipulé (« AI + MODIFIED ») afin d'accroître la transparence pour les personnes exposées au contenu et de remplir les objectifs de divulgation du règlement sur l'IA (AI Act). Les conclusions et la méthodologie de l'étude empirique seront rendues publiques.

Ces icônes sont accessibles au public pour que les Signataires puissent les utiliser conformément aux spécifications de placement, sans qu'il soit nécessaire de les attribuer à la Commission ou au Bureau de l'IA.

Le groupe de travail étudiera l'intégration d'une seconde couche interactive dans le but d'enrichir les icônes statiques ci-dessus avec des données de provenance et des informations sur ce qui a été modifié. Le développement de cette couche interactive sera aligné sur les capacités techniques des déployeurs et des autres acteurs distribuant et/ou vérifiant le contenu. Une icône de l'UE uniquement audio sera également développée.

Figure 1. L'icône de l'UE développée par le Bureau de l'IA pour le contenu entièrement généré par IA, en plusieurs variantes sur un arrière-plan (noir, blanc, noir transparent, blanc transparent).

Figure 2. L'icône de l'UE développée par le Bureau de l'IA pour le contenu partiellement modifié par IA, en plusieurs variantes sur un arrière-plan (noir, blanc, noir transparent, blanc transparent).

Figure 3. L'icône de l'UE de base développée par le Bureau de l'IA en plusieurs variantes sur un arrière-plan (noir, blanc, noir transparent, blanc transparent).