Pratiques responsables de PAI en matière de médias synthétiques

Un cadre pour l'action collective

ATTENTION
Cette traduction Française inclus des termes techniques qui n'ont pas forcément d'équivalents. En cas d'ambiguïté, consultez des experts ou prenez référence sur la version originale en Anglais du Cadre.

FÉVRIER 2023

PARTNERSHIP ON AI

Les pratiques responsables de Partnership on AI en matière de médias synthétiques sont un ensemble de recommandations qui visent à encadrer le développement et le déploiement responsables des médias synthétiques.

Ces pratiques ont été élaborées à partir de commentaires formulés par plus d'une centaine de partenaires mondiaux. Elles sont fondées sur le travail réalisé par PAI ces quatre dernières années avec des représentants de l'industrie, de la société civile, des médias et du journalisme, et du secteur universitaire.

Ce cadre devrait nous permettre de :

1. Mieux comprendre comment les avantages des médias synthétiques peuvent être exploités de façon responsable, en établissant un consensus et une communauté autour des meilleures pratiques, avec la collaboration de partenaires clés de l'industrie, des médias, des milieux universitaires et de la société civile.
2. Offrir à la fois des conseils aux nouveaux acteurs du domaine des médias synthétiques, et une réflexion sur les pratiques d'acteurs plus importants du domaine.
3. Harmoniser les normes et les pratiques pour réduire la redondance et faciliter l'adoption de pratiques responsables dans l'ensemble de l'industrie et de la société, de sorte à éviter un nivellement par le bas.
4. Nous assurer qu'il existe un document et une communauté associée qui sont à la fois utiles et adaptables aux derniers développements dans un environnement naissant et évoluant rapidement.
5. Servir de complément aux autres normes et politiques en matière de médias synthétiques, notamment à l'échelle internationale.

Gouvernance et participation

Les parties intéressées visées par ces pratiques regroupent des créateurs de technologies et d'outils utilisés pour les médias synthétiques, ou les personnes qui créent, partagent et publient ce type de contenu.

Plusieurs de ces parties intéressées seront associées officiellement aux activités entourant le lancement des pratiques responsables de PAI en matière de médias synthétiques. Ces organisations :

1. Participeront à la communauté de pratique de PAI;
2. Contribueront à l'analyse ou l'étude de cas annuelle qui examine l'utilisation du cadre dans les technologies ou les pratiques de produit.

PAI ne procèdera pas à l'audit ou à la certification d'organisations. Le cadre présente les pratiques suggérées qui servent de base à l'orientation.

Le présent document sur les pratiques responsables de Partnership on AI (PAI) est un document évolutif. Même s'il reflète les normes et les pratiques actuelles, il évoluera pour tenir compte des dernières avancées de la technologie, des cas d'utilisation et des nouveaux partenaires. Les médias synthétiques responsables (développement de l'infrastructure, création et diffusion) constituent un phénomène émergent, qui se développe rapidement et nécessite de la souplesse et des recalibrages au fil du temps. PAI prévoit mener un examen annuel du cadre et lancer des examens ponctuels chaque fois que le comité directeur sur l'intégrité de l'IA et des médias l'estimera nécessaire.

L'objectif du cadre

Les médias synthétiques ne sont pas nécessairement dangereux, mais la technologie est de plus en plus accessible et sophistiquée, ce qui amplifie à la fois les risques et les occasions d'utiliser la technologie de façon responsable et même bénéfique, que nous commençons à peine à aborder dans ce cadre. Le cadre vise donc principalement à étudier la manière de gérer

les risques que les médias synthétiques peuvent poser, tout en permettant d'en tirer parti de façon responsable.

De plus, même si les implications éthiques des médias synthétiques sont vastes, puisqu'elles concernent des éléments comme le droit d'auteur, l'avenir du travail, et même la signification de l'art, ce document cible initialement certains groupes de parties intéressées que la communauté de PIA travaillant sur l'intégrité de l'IA et des médias considère comme pouvant jouer un rôle significatif pour : a) réduire les effets néfastes potentiels d'une mauvaise utilisation des médias synthétiques et promouvoir leur utilisation responsable; b) accroître la transparence, et c) donner les moyens aux auditoires de mieux repérer les médias synthétiques et d'y réagir.

Pour en savoir plus sur la création, les objectifs et l'évolution des pratiques responsables de PAI en matière de médias synthétiques, voir la FAQ à ce sujet.

Pratiques responsables de PAI en matière de médias synthétiques

Les personnes qui conçoivent la technologie et l'infrastructure pour les médias synthétiques, qui créent les médias synthétiques, qui les partagent ou qui les publient chercheront à favoriser un comportement éthique et responsable.

Nous définirons ici les médias synthétiques, également appelés médias génératifs, comme un contenu visuel, auditif ou multimodal qui a été généré ou modifié (habituellement au moyen de l'intelligence artificielle). Souvent très réalistes, les contenus produits ne sembleront pas synthétiques pour la plupart des gens; ils peuvent simuler des objets, des personnes ou des événements. Voir l'annexe A pour en apprendre davantage sur la portée du cadre.

PAI formule des **recommandations pour les différentes catégories de parties intéressées** en ce qui concerne leurs rôles dans le développement, la création et la diffusion des médias synthétiques. **Ces catégories ne sont pas mutuellement exclusives.** Une partie intéressée donnée pourrait correspondre à plusieurs catégories, comme c'est le cas pour les plateformes de médias sociaux. Ces catégories incluent :

- Les créateurs de technologies et d'outils utilisés pour les médias synthétiques
- Les créateurs de médias synthétiques
- Les personnes qui partagent et publient des médias synthétiques

SECTION 1

Pratiques pour favoriser l'utilisation éthique et responsable des médias synthétiques

1. Collaborer pour faire progresser la recherche, les solutions techniques, les initiatives d'éducation aux médias et les propositions de politiques pour appuyer la lutte contre les utilisations néfastes des médias synthétiques. Notons que les médias synthétiques peuvent être utilisés de façon responsable ou de façon à causer du tort.

   Les catégories d'utilisation responsable peuvent comprendre les éléments suivants, sans toutefois s'y limiter :

   - Divertissement
   - Arts
   - Satire
   - Éducation
   - Recherche

2. Mener des recherches et faire connaître les pratiques exemplaires afin de poursuivre la définition des catégories d'utilisation responsables ou néfastes des médias synthétiques.

   Ces utilisations comportent souvent des zones grises, et les techniques pour trouver des repères dans ce contexte sont décrites dans les sections ci-dessous.

3. Quand les techniques ci-dessous sont utilisées pour créer ou diffuser du contenu produit avec les médias synthétiques dans le but de causer du tort (voir l'annexe B), adopter des mesures d'atténuation raisonnables, en conformité avec les méthodes décrites dans les sections 2, 3 et 4.

Les techniques suivantes peuvent être déployées tant de façon responsable que pour causer du tort :

- Représenter une personne ou une entreprise, une organisation médiatique, un organisme gouvernemental, ou une entité;
- Créer de faux personnages réalistes;
- Représenter une personne en particulier comme agissant, se comportant ou faisant des déclarations d'une manière qui ne correspond pas à cette personne;
- Représenter des événements ou des interactions qui n'ont jamais eu lieu;
- Insérer des éléments créés synthétiquement ou supprimer des éléments authentiques dans du véritable contenu média;
- Générer des scènes ou des environnements sonores entièrement synthétiques.

Vous trouverez des exemples de la manière dont ces techniques peuvent être utilisées pour causer du tort, ainsi qu'une liste non exhaustive de leurs effets néfastes, dans l'annexe B.

SECTION 2

Pratiques pour les **concepteurs de technologies et d'infrastructure**

Les personnes qui conçoivent et fournissent les technologies et l'infrastructure utilisées pour les médias synthétiques peuvent comprendre : les concepteurs d'outils pour le commerce interentreprises (B2B) et le commerce grand public (B2C); les développeurs de logiciels ouverts; les chercheurs universitaires; les jeunes entreprises à l'œuvre dans les médias synthétiques, notamment celles fournissant aux amateurs l'infrastructure nécessaire pour créer des médias synthétiques; les plateformes de médias sociaux; et les boutiques d'applications.

4. **Faire preuve de transparence envers les utilisateurs au sujet de ces outils et technologies**, de leurs capacités, leurs fonctionnalités et leurs limites, ainsi que des risques posés par les médias synthétiques.

5. Prendre des mesures pour **fournir des mécanismes de divulgation** aux créateurs et aux distributeurs de médias synthétiques.

   La divulgation peut se faire **directement ou indirectement**, en fonction des cas d'utilisation et du contexte :

   - La divulgation directe vise directement le téléspectateur ou l'auditeur et peut comprendre, sans toutefois s'y limiter, des étiquettes de contenu, des notes contextuelles, des tatouages numériques et des avertissements.
   - La divulgation indirecte est intégrée et comprend, sans toutefois s'y limiter, la provenance cryptographique des contenus synthétiques (comme la norme C2PA), et l'application d'éléments traçables aux données d'apprentissage et aux résultats, aux métadonnées du fichier de média synthétique, à la composition des pixels, et à des énoncés de divulgation sur une image dans les vidéos.

6. **Lors du développement de codes et d'ensemble de données, des modèles de formation et des logiciels qui s'appliquent** à la production de contenus médias synthétiques, **faire son possible pour appliquer des éléments de divulgation indirecte** (stéganographie, origine du contenu média ou autres) aux différents actifs ainsi qu'aux différentes étapes de la production de contenu média synthétique.

   L'objectif consiste à divulguer l'information de manière à limiter les suppositions quant au contenu, à faire le maximum en matière de résistance à la manipulation et à la falsification, à être précis, et aussi, s'il y a lieu, à communiquer l'incertitude sans ajouter aux spéculations.

7. **Soutenir la recherche pour orienter les initiatives de partage de données à venir** (accroître la transparence tout en respectant la protection des renseignements personnels), et déterminer quels types de données seraient les plus appropriées et les plus intéressantes à recueillir et à communiquer.

8. Prendre des mesures pour **effectuer de la recherche et procéder au développement et au déploiement** de technologies qui :

   - Permettent le plus possible de détecter les manipulations sans freiner l'innovation en matière de photoréalisme;
   - Permettent la divulgation durable du contenu synthétique au moyen de méthodes comme le tatouage électronique ou la cryptographie de l'origine du contenu, qui sont découvrables, protègent les renseignements personnels, sont facilement accessibles pour la communauté étendue, et proviennent de codes sources ouverts.

9. **Prévoir une politique publiée et accessible** sur l'utilisation éthique de ces technologies, mentionnant les restrictions d'utilisation que les utilisateurs auront à respecter et que les fournisseurs devront appliquer.

SECTION 3
Pratiques pour les **créateurs**

Ces pratiques s'appliquent aux créateurs de médias synthétiques, des producteurs à grande échelle (comme les producteurs de contenu interentreprises) aux plus petits producteurs (comme les amateurs, les artistes, les influenceurs et d'autres dans la société civile, notamment les militants et les auteurs de satire). Elles s'appliquent aussi aux personnes qui commandent ou dirigent le travail de création de contenus médias synthétiques. Compte tenu de la nature de plus en plus démocratisée de ces outils, tout le monde peut devenir un créateur et avoir la possibilité que le contenu produit touche un large public. Ces différentes parties intéressées sont par conséquent données à titre illustratif et leur liste n'est pas exhaustive.

10. **Faire preuve de transparence** envers les consommateurs de contenu pour ce qui suit :

    - La manière dont vous avez obtenu le **consentement éclairé** de la part du ou des sujets présents dans du contenu manipulé, en fonction du produit et du contexte, sauf si le contenu est utilisé raisonnablement à des fins artistiques, satiriques ou expressives.
    - La manière dont vous percevez l'utilisation éthique de la technologie et les restrictions d'utilisation (p. ex. au moyen d'une politique **publiée** et accessible sur votre site web ou dans des publications sur votre travail) et consultez ces lignes directrices avant de créer du contenu média synthétique.
    - Les capacités, les limites et les risques associés au contenu synthétique.

11. **Procéder à la divulgation** lorsque le contenu média que vous avez créé ou intégré comporte des éléments synthétiques, particulièrement lorsque le fait d'ignorer la présence de ces éléments synthétiques modifie la manière dont le contenu est perçu. Utiliser tous les outils de divulgation fournis par les concepteurs des technologies et de l'infrastructureservant à créer du contenu de médias synthétiques.

    La divulgation peut se faire **directement ou indirectement**, en fonction des cas d'utilisation et du contexte :

    - La divulgation directe vise directement le téléspectateur ou l'auditeur et peut comprendre, sans toutefois s'y limiter, des étiquettes de contenu, des notes contextuelles, des tatouages numériques et des avertissements.

- La divulgation indirecte est intégrée et comprend, sans toutefois s'y limiter, la provenance cryptographique des contenus synthétiques (comme la norme ouverte C2PA), et l'application d'éléments traçables aux données d'apprentissage et aux résultats, aux métadonnées du fichier de média synthétique, à la composition des pixels, et à des énoncés de divulgation sur une image dans les vidéos.

L'objectif consiste à divulguer l'information de manière à limiter les suppositions quant au contenu, à faire le maximum en matière de résistance à la manipulation et à la falsification, à être précis, et aussi, s'il y a lieu, à communiquer l'incertitude sans ajouter aux spéculations.

SECTION 4

Pratiques pour les **distributeurs et les éditeurs**

Les distributeurs de médias synthétiques comprennent à la fois les entreprises qui produisent du contenu actif faisant l'objet de décisions de nature éditoriale et qui hébergent principalement du contenu propriétaire (comme les entreprises médiatiques, notamment les radiodiffuseurs), ainsi que des plateformes web dont le contenu synthétique est plus passif et généré par des utilisateurs ou des tiers (comme les plateformes de médias sociaux). C'est le cas d'organisations médiatiques qui distribuent du contenu média synthétique créé à des fins éditoriales, ou qui produisent de l'information sur le contenu média synthétique créé par d'autres.

Pour les canaux de distribution active et passive

12. **Procéder à la divulgation** lorsque vous avez déterminé avec certitude la présence de contenu synthétique généré par des tiers ou des utilisateurs.

    La divulgation peut se faire **directement ou indirectement**, en fonction des cas d'utilisation et du contexte :

    - La divulgation directe vise directement le téléspectateur ou l'auditeur et peut comprendre, sans toutefois s'y limiter, des étiquettes de contenu, des notes contextuelles, des tatouages numériques et des avertissements.
    - La divulgation indirecte est intégrée et comprend, sans toutefois s'y limiter, la provenance cryptographique des contenus synthétiques (comme la norme ouverte C2PA), et l'application d'éléments traçables aux données de formation et aux résultats, aux métadonnées du fichier de média synthétique, à la composition des pixels, et à des énoncés de divulgation sur une image dans les vidéos.

    L'objectif consiste à divulguer l'information de manière à limiter les suppositions quant au contenu, à faire le maximum en matière de résistance à la manipulation et à la falsification, à être précis, et aussi, s'il y a lieu, à communiquer l'incertitude sans ajouter aux spéculations.

13. **Afficher une politique** publiée et accessible qui énonce l'approche adoptée et appliquée par l'organisation en matière de médias synthétiques.

Pour les canaux de distribution active

Les canaux (p. ex. organisations médiatiques) qui hébergent principalement du contenu propriétaire et qui peuvent distribuer ou signaler du contenu de médias synthétiques de nature éditoriale créé par d'autres.

14. **Apporter rapidement des ajustements** quand on se rend compte que du contenu synthétique néfaste a été distribué ou est présent sans qu'on le sache.

15. **Éviter de distribuer** du **contenu** média synthétique **sans en citer la source** ou d'inclure dans des reportages du contenu média synthétique néfaste créé par d'autres sans avertissement clair ni contexte, afin de s'assurer que tout téléspectateur ou lecteur raisonnable comprendra qu'il s'agit de contenu synthétique.
16. **S'orienter vers une** infrastructure organisationnelle qui permet de déterminer la **provenance du contenu**, tant pour les médias synthétiques que non synthétiques, tout en respectant la confidentialité (par exemple au moyen de la norme ouverte C2PA).
17. **S'assurer qu'un consentement transparent et éclairé** a été fourni par **le créateur et le ou les sujets représentés** dans le contenu synthétique qui sera communiqué ou distribué, même si on a déjà obtenu le consentement pour la création du contenu.

Pour les canaux de distribution passive

Les canaux (p. ex. plateformes) qui hébergent principalement du contenu tiers.

18. **Mettre en œuvre** des **mesures raisonnables de signalement** sur les aspects techniques, les utilisateurs et le personnel afin de signaler la présence de contenu néfaste distribué sur les plateformes.
19. **Apporter rapidement des ajustements** au moyen d'étiquettes, de déclassement ou de suppression du contenu, ou d'autres interventions comme celles décrites ici, lorsqu'on sait que du contenu synthétique néfaste est distribué sur des plateformes.
20. **Informer** les utilisateurs des plateformes et les **sensibiliser** au sujet des médias synthétiques et des types de contenus synthétiques admissibles créés et publiés sur les plateformes.

Annexes

ANNEXE A

Portée des pratiques responsables de Partnership on AI (PAI) en matière de médias synthétiques

Même si ce document se concentre sur les formes de médias synthétiques les plus réalistes, il reconnaît que le seuil de ce qui est jugé comme très réaliste peut varier selon le niveau d'éducation aux médias des auditoires, et en fonction des différents contextes mondiaux. Nous reconnaissons aussi que les médias synthétiques moins réalistes peuvent quand même avoir des effets néfastes, comme dans le cas d'utilisation frauduleuse d'images intimes.

Le cadre vise principalement les médias synthétiques audiovisuels, également appelés médias génératifs, plutôt que les textes synthétiques, qui présentent des avantages et des risques différents. Cependant, le cadre offre aussi des conseils utiles sur la création et la distribution de textes synthétiques.

Par ailleurs, le cadre ne couvre que les médias génératifs, et non la catégorie plus large de l'IA générative dans son ensemble. Nous sommes conscients que ces termes sont parfois utilisés de manière interchangeable.

Les médias synthétiques ne sont pas nécessairement dangereux, mais la technologie est de plus en plus accessible et sophistiquée, ce qui amplifie à la fois les risques et les occasions. Nous réviserons le cadre en fonction des avancées technologiques et l'adapterons aux changements technologiques (par exemple les expériences de médias immersifs).

ANNEXE B

Risques associés aux médias synthétiques

Voici une liste des risques présentés par les médias synthétiques qu'il faut chercher à réduire :

- Usurpation d'identité pour obtenir de l'information confidentielle ou des privilèges.
- Appels téléphoniques, communications de masse, publications ou messages destinés à tromper ou à harceler.
- Fraudes visant l'obtention d'avantages financiers.
- Désinformation au sujet d'une personne, d'un groupe ou d'une organisation.
- Exploitation ou manipulation d'enfants.
- Intimidation et harcèlement.
- Espionnage.
- Manipulation de processus démocratiques et politiques, notamment : duperie à l'endroit d'électeurs pour qu'ils votent en faveur ou contre un candidat, atteinte à la réputation d'un candidat au moyen d'affirmations ou d'actes falsifiés, et influence du résultat d'une élection par la tromperie ou la suppression de vote.
- Manipulation des marchés et sabotage d'entreprises.
- Création de discours haineux, de discrimination, de diffamation, de terrorisme ou d'actes violents, ou incitation à en créer.
- Diffamation et sabotage de réputation.
- Diffusion non consensuelle de contenu intime ou sexuel.
- Extorsion et chantage.
- Création en masse de nouvelles identités et de nouveaux comptes représentant prétendument des personnes distinctes de manière à « fabriquer de l'opinion publique ».