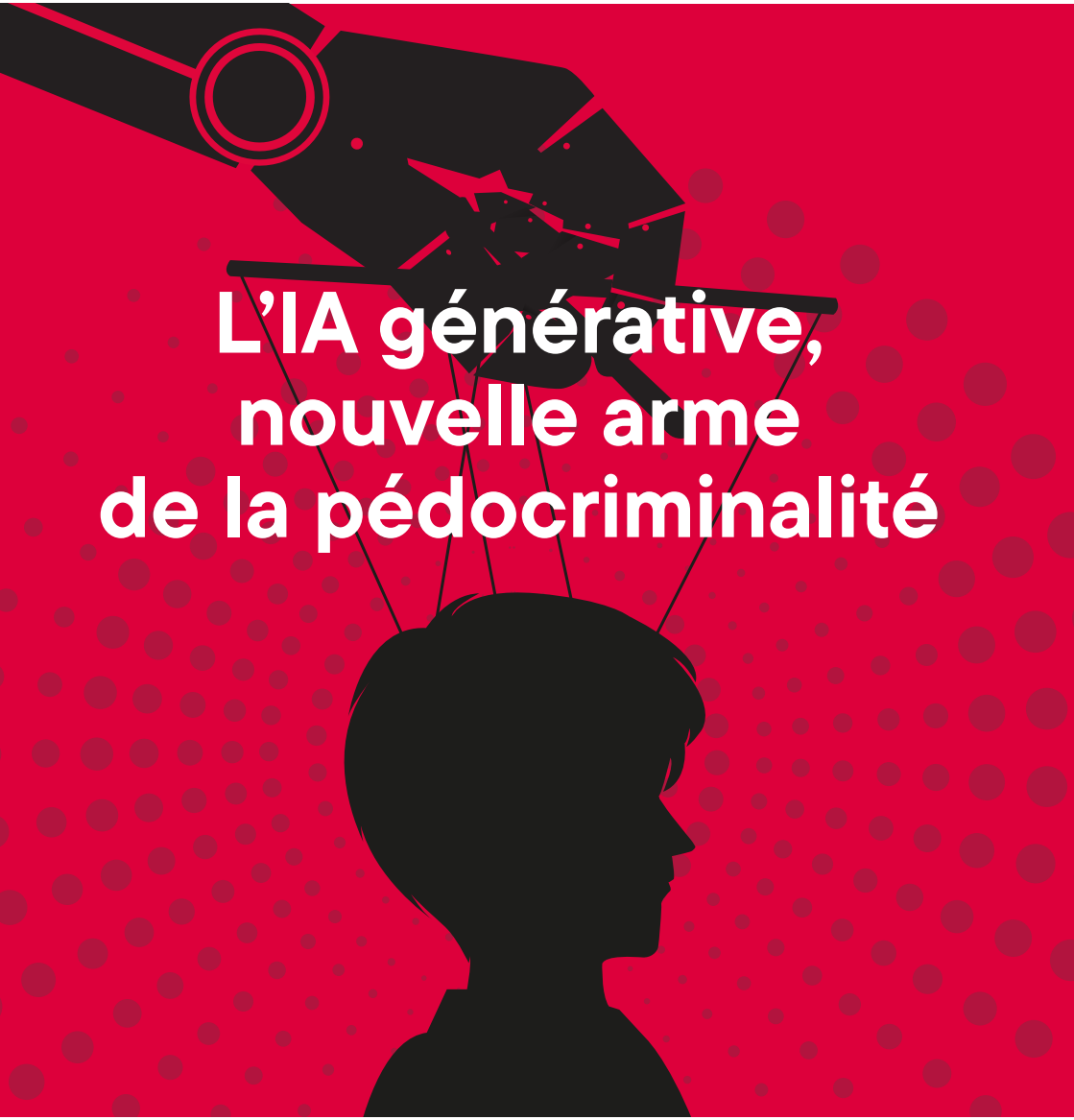
Synthèse du rapport de la **Fondation pour l'Enfance**

L'IA générative, nouvelle arme de la pédocriminalité



Octobre 2024

FONDATION POUR L'ENFANCE
reconnue d'utilité publique

Si elle représente une évolution technologique majeure, l'intelligence artificielle générative se transforme en une arme redoutable lorsqu'elle est utilisée par des personnes mal intentionnées, comme les cyberpédocriminels. Sextorsion, grooming, vidéos pédocriminelles[1]… Ces pratiques sont désormais facilitées et amplifiées par cette technologie qui ne cesse de se perfectionner.

Elles occasionnent notamment des difficultés pour les forces de l'ordre, qui peinent à distinguer les images non générées par l'IA de celles générées par l'IA, et donc à identifier les enfants victimes de violences. Sans compter l'inadéquation du cadre juridique et de la législation, qui donne aux cyberpédocriminels un sentiment d'impunité et de toute puissance. Le caractère virtuel de ces montages entraîne chez leurs auteurs et consommateurs une tendance à la banalisation et à la normalisation de ces pratiques. Pourtant, c'est l'enfant lui-même, son intégrité physique et morale, ses droits qui sont attaqués.

Bien que la part de ces contenus dans l'ensemble des signalements soit encore faible, et ce au niveau mondial, la Fondation pour l'Enfance et ses partenaires tirent la sonnette d'alarme. Après avoir mené une recherche approfondie pendant près d'un an, leur constat est sans appel : la partie immergée de ces contenus pédocriminels dopés à l'IA semble être colossale, et leurs conséquences pour les victimes, dramatiques. Il est donc urgent d'engager une réponse forte, rapide et coordonnée entre les différents acteurs juridiques, politiques, technologiques, afin d'apporter un cadre légal à l'utilisation de l'IA, mais aussi de permettre une prise de conscience sociétale autour des risques liés à l'IA et du danger lié à certaines pratiques, notamment du *sharenting*.

Nos objectifs ? Protéger et assurer la sécurité et l'intégrité en ligne des enfants.

1. Un glossaire expliquant certains termes liés à la cyberpédocriminalité et à l'IA générative se trouve en fin de rapport

Pédocriminalité & IA : nouveaux usages, nouveaux enjeux, nouveaux risques

L'essor de l'IA générative nous confronte à cette question : comment distinguer les contenus (textes, images, audios, vidéos) créés *ex nihilo* de ceux qui montrent des personnes réelles ? Sur le terrain de la lutte contre la cybercriminalité, cette question prend une ampleur vertigineuse.
En cause : l'émergence de nouveaux usages pédocriminels complexes et dangereux.

Les nouveaux usages pédocriminels avec l'IA générative

Des vidéos d'enfants qui n'existent pas en train d'être violés, des visages de vraies adolescentes dont le corps est « synthétiquement » entièrement dénudé… L'IA générative permet de générer à l'infini de tels contenus, brouillant ainsi les pistes entre réalité et virtuel. En somme, un véritable « terrain de jeu » pour cyberpédocriminels, dont voici quelques-uns des nouveaux usages.

Modification de modèles et systèmes d'IA, dans le but de les faire générer des contenus pédocriminels :

- Certains modèles et systèmes d'IA *open source* et disponibles au grand public sont modifiés par des individus, dans le but de créer des contenus pédocriminels
- Ces modèles et systèmes modifiés sont entraînés sur de larges bibliothèques de contenus d'exploitation sexuelle de mineurs, de contenus pornographiques, mais aussi sur des images d'enfants à caractère non sexuel, afin de leur enseigner comment produire avec précision du nouveau matériel pédocriminel.

Ces contenus sur lesquels les modèles sont entraînés sont des images et vidéos, présents en grande quantité sur internet et facilement accessibles.

- La modification des modèles et la génération d'images pédocriminelles peuvent être effectuées hors ligne sur un ordinateur personnel, permettant ainsi d'échapper à la détection.

Création et modification de contenus démultipliés par les IA sur le *dark web* et le *clear web*:

- La création de contenus pédocriminels *ex nihilo*: des images, photos et vidéos pédocriminelles, totalement artificielles, ressemblant à s'y méprendre à des agressions sexuelles/viols réels.

- La création de *deepfakes* pédocriminels: des images, photos et vidéos pédocriminelles, générées à partir d'autres contenus d'enfants à caractère sexuel, ou même non-sexuel, présents sur internet (comme les réseaux sociaux). Ces montages sont générés à partir de modèles d'IA modifiés ou de *nudify apps*, des applications IA de déshabillage.

- L'édition et l'amélioration de la qualité de photos et vidéos pédocriminelles déjà existantes.

- Des instructions données aux modèles d'IA générative pour créer et affiner des guides et des tutoriels sur la manière de gagner la confiance, de violer, d'agresser, de torturer et de tuer des enfants, ou de créer des contenus pédocriminels réalistes.

LE POINT DE VUE DES EXPERTES

JOANNA SMITH, psychologue clinicienne et **MÉLANIE DUPONT,** docteur en psychologie, psychologue à l'Unité Médico-Judiciaire de l'Hôtel-Dieu (Paris) et présidente de l'Association contre les Violences sur Mineurs (CVM)

Quels impacts de la cyberpédocriminalité pour les victimes ?

Il est difficile, voire impossible pour la victime de voir une fin à l'épisode traumatique, parce que le contenu continue de tourner sur internet. Cette fin est pourtant cruciale dans le traitement du traumatisme. Sans celle-ci, le danger est encore présent. Avec la cyberpédocriminalité, il y a une permanence de l'agression, de multiples agresseurs (ceux qui visionnent, téléchargent, partagent le contenu), et donc une permanence des conséquences et une revictimisation. La création de contenus pédocriminels par l'IA générative, et donc l'impossibilité de contrôler son image, va entraîner chez les victimes une intensification du sentiment de dépossession de soi. Cela pourrait avoir pour effet d'augmenter les troubles psychotiques chez les jeunes, et même une dépersonnalisation à se voir sur des images qui ne représentent pas leurs propres corps.

Le business de la cyberpédocriminalité dopé à l'IA

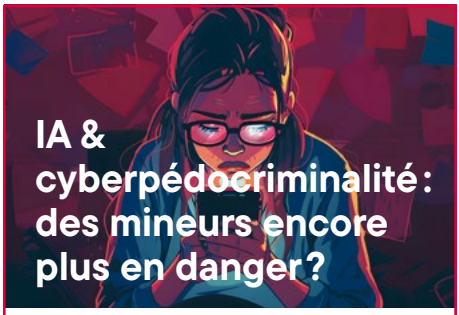
Comme à chaque avancée technologique, l'innovation se transforme en consommation. La cyberpédocriminalité via IA générative ne fait pas exception à la règle : un nouveau business est en train d'émerger.

- Pour publier, partager et/ou faire la promotion de leurs contenus, les pédocriminels utilisent le *clear web* avec de faux profils sur les réseaux sociaux (Instagram, Facebook, Tik Tok) et les plateformes de messagerie cryptées (Telegram, Whatsapp).

- Certains profils encouragent les abonnés à contacter le propriétaire via le système de messagerie de la plateforme, ou via une plateforme chiffrée tierce, pour obtenir plus d'images, toujours plus graphiques et explicites. Nombre d'entre eux mettent leurs contenus à disposition moyennant paiement, et relaient leurs « clients » vers des systèmes de paiement et autres services d'abonnement (tels qu'OnlyFans).

- Si un utilisateur ne parvient pas à générer ce qu'il souhaite, ou si un modèle ou un fichier n'existe pas encore, il peut être amené à payer un utilisateur plus qualifié sur l'IA pour le faire à sa place.

- Certains pédocriminels font du chantage à des mineurs pour obtenir de leur part de l'argent ou des contenus à caractère sexuel (sextorsion). Cette pratique se fait à l'aide de montages pédocriminels générés avec l'IA, à partir de photos à caractère non-sexuel, obtenues notamment sur les réseaux sociaux.



IA & cyberpédocriminalité : des mineurs encore plus en danger ?

52 % des consommateurs pensent que leur usage de contenus pédocriminels pourrait aboutir à une agression sur un enfant (44 % des consommateurs ont pensé à contacter des enfants et 37 % ont contacté des enfants au moins une fois)*.

Puisque l'IA générative permet une création de contenus à l'infini, elle augmente les comportements addictifs des consommateurs, avec des images de plus en plus extrêmes, explicites, violentes. Et donc, des risques accrus de passage à l'acte.

*Protect Children, *ReDirection Survey Report,* 2021, p.16

Les 3 principaux enjeux pour protéger les mineurs de la cyberpédocriminalité générée par l'IA

L'émergence des contenus pédocriminels générés par l'IA amplifie les enjeux déjà existants en matière de protection des enfants, mais en crée également de nouveaux:

Identifier et protéger les enfants victimes

Par leur réalisme bluffant, les contenus pédocriminels générés par IA rendent la tâche d'identification et de protection des enfants victimes de violences sexuelles difficile pour les forces de l'ordre et les plateformes de signalement.

Engager une réponse juridique et politique solide et coordonnée

À l'heure actuelle, il n'existe pas de législation qui traite avec

Nos recommandations

La Fondation pour l'Enfance appelle les États et les entreprises du secteur des nouvelles technologies à agir au plus vite. Trois objectifs impératifs sont à atteindre:

Prévention

Mettre en place des campagnes nationales de sensibilisation du grand public à la cyberpédocriminalité, aux dangers relatifs au *sharenting* et aux bonnes pratiques à adopter pour protéger les enfants. Cette campagne doit permettre aux parents d'appréhender leur rôle en matière de prévention et d'être davantage sensibilisés aux risques liés à l'IA.

Détection

Favoriser l'innovation, en incitant les acteurs privés à coopérer pour mettre en place des outils permettant de distinguer les contenus générés par l'IA des contenus non générés par l'IA. L'utilisation de tels outils permettrait de pallier la difficulté d'identification des mineurs victimes de violences.

Instaurer un travail conjoint entre les différentes entreprises et les plateformes, afin d'améliorer l'identification, le signalement et le retrait des contenus pédocriminels et des modèles d'IA générative destinés à les générer.

Sanction

Amender l'article 227-23 du Code pénal pour y insérer les fichiers ou représentations issus de l'IA avec l'insertion d'un nouvel alinéa qui pourrait être rédigé comme suit: *« Le fait de concevoir, de créer, de diffuser ou de porter à la connaissance du public ou d'un tiers, par quelque voie que ce soit, tout montage, contenu visuel ou sonore à caractère*

précision de la création, de la possession ou du partage d'un modèle d'IA générative conçu pour produire des contenus pédocriminels.

Ce vide juridique et cette absence de législation nationale et internationale claire, assortie d'une absence de prise de conscience sociétale, permet à la pratique de s'intensifier.

Lutter contre l'intensification et la banalisation des violences sexuelles sur les enfants

La facilité de production des contenus pédocriminels générés par l'IA et leur multiplication sur Internet entraînent une normalisation et une banalisation des violences sexuelles sur les enfants.

sexuel généré par un traitement algorithmique tel que visé à l'alinéa 1 de l'article 226-8-1 est puni de X ans de prison et X euros d'amende lorsqu'il s'agit de la représentation, de l'image ou de la parole d'un mineur. »

Pénaliser la création et la mise à disposition de modèles d'IA générative destinés à générer des contenus pédocriminels. Il pourrait être ajouté au Code pénal un nouvel article rédigé de la manière suivante : « *Est puni de X années d'emprisonnement et X euros d'amende le fait de collecter, détenir, traiter ou détourner des données à caractère personnel, afin de créer, générer ou mettre à disposition du public ou de tout tiers un modèle de traitement algorithmique, dans le but de permettre la création de contenu visuel ou sonore à caractère sexuel représentant un mineur, et de tout fichier à caractère pédopornographique.* »

Quand les géants de la tech' s'engagent

Meta, Google, Microsoft, OpenAI, Amazon... Les leaders de l'IA ont adopté le 23 avril 2024 le texte "Safety by Design for Generative AI: Preventing Child Sexual Abuse" proposé par l'organisation à but non lucratif Thorn. Au cœur de ces engagements ? La mise en place de nouvelles mesures de sécurité pour protéger les enfants en ligne, et une meilleure prise en compte de la protection des enfants dans le développement et le déploiement de l'IA générative, afin d'empêcher que ces outils permettent la création de contenus pédocriminels. Si le texte n'a valeur que morale et non juridique, il reste un premier pas (très) inspirant !

Présentation

La Fondation pour l'Enfance est née en 2012 de la fusion de la Fondation pour l'Enfance, fondée en 1977 par Madame Anne-Aymone Giscard d'Estaing, alors Première Dame, et de la Fondation Protection de l'Enfance.

Reconnue d'utilité publique, indépendante et non-partisane, la Fondation pour l'Enfance agit pour améliorer la protection des enfants et le respect de leurs droits fondamentaux, en luttant contre toutes les formes de violences et de maltraitance, et en favorisant des liens adultes-enfants de qualité. Tous ses positionnements et ses recommandations sont validés avec des experts (médecins, sociologues, psychologues, professionnels de la petite enfance, avocats etc.)

La Fondation pour l'Enfance intervient auprès de l'ensemble des acteurs institutionnels, associatifs et privés qui agissent dans le secteur de l'enfance. Elle agit à travers des actions de plaidoyer auprès des pouvoirs publics (en propre ou via des collectifs interassociatifs) et des actions de sensibilisation et de prévention auprès du grand public et des professionnels (médecins, assistant.e.s maternel.le.s, sage-femmes etc.).

FONDATION POUR L'ENFANCE
reconnue d'utilité publique

Nos partenaires

Valeuriad

Lombard Baratelli Astolfe & associés

aadh
Alliance des Avocats pour les droits de l'Homme

Retrouvez l'actualité de la Fondation pour l'Enfance

FONDATION POUR L'ENFANCE
23, place Victor Hugo, 94 270 Kremlin-Bicêtre

01 43 90 63 10
fondation-enfance.org

Contact
Angèle Lefranc, chargée de plaidoyer
angele.lefranc@fondation-enfance.org

Conception graphique : Agence Panteo.fr/x.jacobi@panteo.fr • Illustration : Adobe Stock