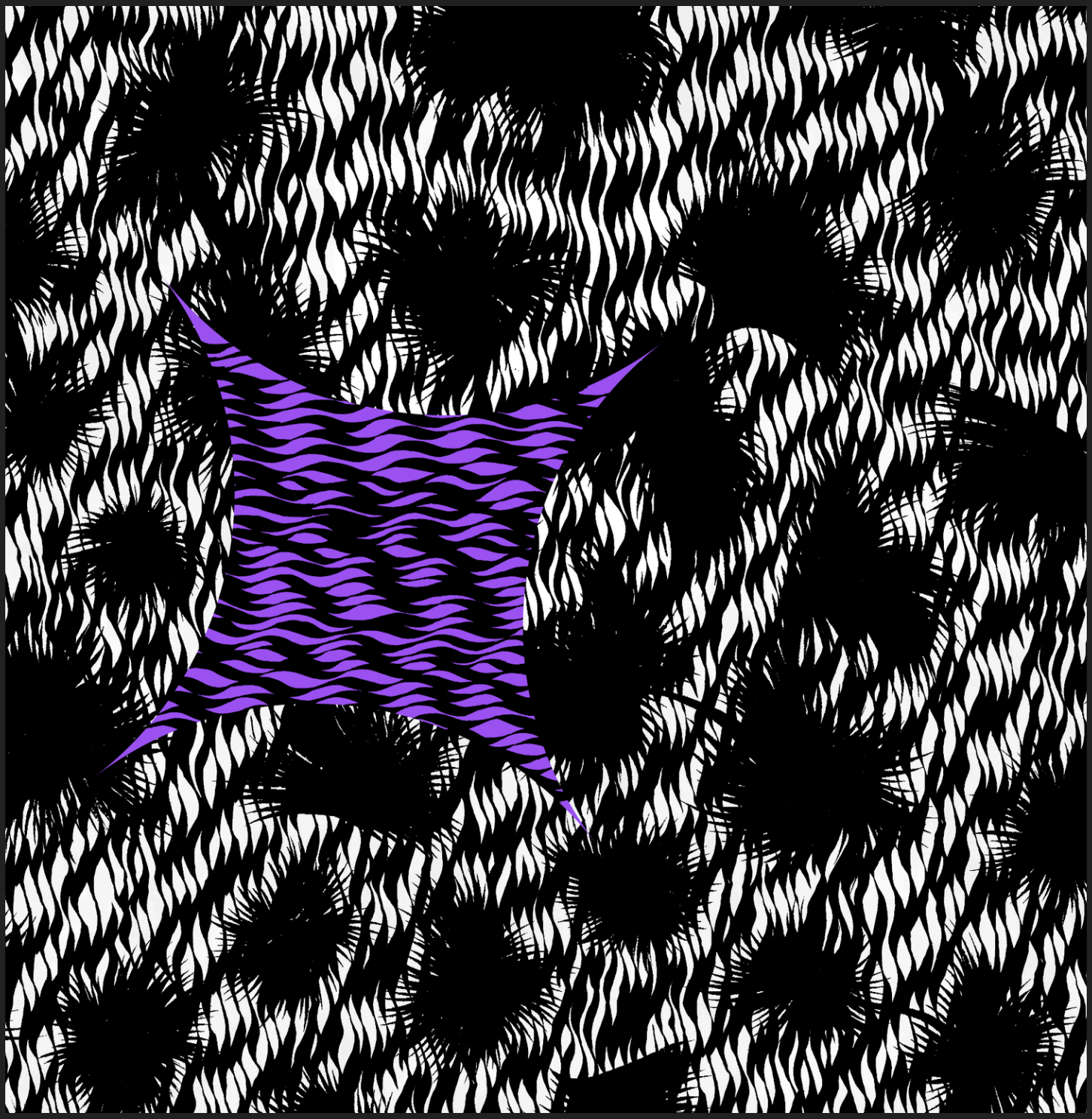
UN PROJET D'ÉDUCATION AUX MÉDIAS QUI QUESTIONNE L'ÉTHIQUE ET L'INTELLIGENCE ARTIFICIELLE

OK MILA

N°04 · IA ET CYBER-HARCÈLEMENT



Editrice, rédactrice et co-fondatrice du projet

Cassi Ninja (Tapage Studio)

Co-fondatrice du projet

Alyssia Ricci

Mise en page du mag papier

Gaëlle Defeyt (Gilda Fêlée)

Mise en page mag online

Cassi Ninja

Charte graphique

François d'Alcamo

Communication et réseaux sociaux

Anne-Sophie Skit

Journaliste podcast et voix off

Julie Mouvet

Réalisateur et monteur vidéo

Francisco Luzemo (Hiola Films)

Réalisateur et monteur podcast

Marius Adam (MadSound)

Relectrice

Elisabeth Bois d'Enghien

WWW.OK-MILA-EAM.BE

OK MILA

CE PROJET EST RENDU POSSIBLE GRÂCE AU CSEM, LE CONSEIL SUPÉRIEUR DE L'EDUCATION AUX MÉDIAS

OK MILA A ÉTÉ CRÉÉ PAR TAPAGE STUDIO-
HELLO@TAPAGE.STUDIO - WWW.TAPAGE.STUDIO

EDITO Rashka Barbare

IA et cyberharcèlement : patriarcat 2.0

On nous avait promis que l'intelligence artificielle allait changer le monde, elle l'a fait, mais pas comme on s'y attendait, ni comme on le voulait.

Ce nouveau monde livre, pour les femmes et les minorités de genre, de nouveaux dangers. Désormais, les insultes, les menaces, l'humiliation, ou encore le harcèlement n'ont plus seulement un visage humain, mais peuvent maintenant prendre la forme de voix synthétiques, de visages générés, de corps créés pixel par pixel.

L'IA, qui s'est nourrie depuis sa création de contenus sexistes, racistes, homophobes, transphobes... est devenue une nouvelle alliée du patriarcat, silencieuse et sans scrupule. Les deepfakes pornos non consentis se multiplient plus vite qu'un virus, les bots harceleurs attaquent telle une meute désireuse d'écharper toutes celles et ceux qui sortent trop des cadres pour les faire taire et disparaître à jamais. Et les algorithmes de modération ferment les yeux sur la haine pullulante, tout en sanctionnant les victimes qui osent répondre et se défendre.

J'aimerais vous dire que c'est un scénario de SF, que XANA a été réveillé et qu'Aelita va venir entrer le code LYOKO, suivis d'un « retour vers le passé » pour mettre fin à ce cauchemar... (les vrai·es auront la réf). Mais tout ça se passe aujourd'hui dans notre monde ! Chaque minute qui passe, quelque part sur la planète, une femme se fait humilier publiquement par une image ou une vidéo qu'elle n'a jamais prise, jamais tournée... Et chaque seconde, un algorithme estime que sa plainte n'est pas prioritaire, pas légitime.

Il est temps de réclamer des IA programmées par des équipes inclusives, diversifiées et éthiques, de créer des gardes-fous qui protègent les victimes et punissent les coupables, et non pas l'inverse.

L'IA est pour l'instant un outil qui sert et crée un « patriarcat augmenté », qui aide ceux qui veulent nous voir disparaître des espaces en ligne, qui veulent nous empêcher de nous exprimer. Or, elle pourrait devenir un outil d'émancipation si nous la reprenions en main dès à présent, elle pourrait identifier plus vite les harceleurs, archiver les preuves, et bloquer les contenus problématiques AVANT leur diffusion.

L'IA ne doit pas être un témoin muet ni un complice, elle doit devenir une alliée !

Rashka, c'est une meuf incroyable qu'on a rencontré dans le cadre d'un autre projet (un podcast sur les masculinités dans le jeu vidéo. Ça s'appelle "En direct de la cuisine" et ça s'écoute comme du petit lait. Voilà, promotime finito). Rashka le dit super bien dans son edito : le cyberharcèlement, c'est déjà un sujet méga touchy parce qu'il touche à l'intime, à l'identité, aux rapports de pouvoir, et qu'il laisse souvent des cicatrices invisibles. Mais quand on y associe l'intelligence artificielle, on entre dans un terrain encore plus glissant : celui où les violences prennent de nouvelles formes, plus rapides, plus massives, plus difficiles à arrêter.

C'est là que ce numéro 4 d'OK MILA entre en scène. L'idée : comprendre brièvement ce que recouvrent les cyberviolences, sous toutes leurs formes, avec un focus sur les deepnudes, en identifiant comment l'IA en modifie les contours, parfois en les aggravant, parfois en offrant des pistes pour mieux les détecter. Et surtout, explorer si, et comment, elle pourrait devenir une alliée réelle pour protéger plutôt que pour nuire... menu alléchant hein ?

LES CYBER-VIOLENCES, C'EST QUOI ?

Les cyberviolences, c'est tout ce qui relève d'un comportement violent via les outils numériques. Autrement dit, c'est quand la violence passe par un écran, un clavier, un micro, une image ou un message. Ça peut être une insulte publique sur Insta, un commentaire humiliant sur une photo, un vocal menaçant sur WhatsApp, une rumeur lancée dans groupe Snap ... Mais aussi une image volée, modifiée, partagée sans le consentement. Et comme les violences "classiques", les cyberviolences peuvent être psychologiques, sexuelles, sociales, économiques ou tout ça en même temps.

Il existe plein de formes de cyberviolences, plus ou moins connues, plus ou moins visibles.

- Le **cyberharcèlement**, c'est quand une personne (ou un groupe) harcèle en ligne : insultes répétées, attaques personnelles, dénigrements publics.
- Le **sextorsion**, c'est le chantage financier à l'image intime : de faux comptes (catfishs) invitent à s'envoyer des photos ou vidéos intimes et finissent par les menacer de diffuser les contenus si les victimes ne leur envoient pas une somme colossale d'argent.
- Le **doxxing**, c'est la diffusion d'infos personnelles (adresse, numéro, nom de famille...) pour exposer à des dangers réels.
- Le **slut-shaming**, c'est le fait d'humilier ou de rabaisser une fille parce qu'elle est jugée "trop sexy", "trop ouverte" ("trop" tout court en fait...)
- Le **revenge porn**, c'est la diffusion de contenus sexuels, souvent pris ou reçus dans un contexte intime, mais balancés publiquement après une rupture.
- Et aujourd'hui, on parle aussi de **deepfake sexuel, de deepnudes** : des montages pornographiques créés avec l'aide de l'IA, qui montrent une personne (souvent une femme) dans une scène qui n'a jamais eu lieu, mais qui semble vraie (et ce sera le sujet principal de ce numéro).

Tous ces actes peuvent avoir des conséquences très graves, même s'ils se déroulent "juste" en ligne (oui oui les boomers, je vous entends) : anxiété, isolement, peur, repli, parfois décrochage scolaire ou perte d'emploi. Et non, il ne suffit pas de "couper internet" ou de "ne pas répondre" pour que ça s'arrête. Car les cyberviolences, ce sont des violences, et il faut remettre la responsabilité sur les personnes qui causent ces violences. C'est tout. (Ça fait beaucoup de fois le mot "violence" mais c'est le thème, vous m'excuserez).

QUI EST CONCERNÉ?

On peut considérer les cyberviolences, comme une extension des violences systémiques, version numérique. Ce sont souvent les mêmes personnes qui sont déjà ciblées hors ligne : les femmes, les personnes queer, les jeunes, les personnes non blanches, ou les personnes en situation de handicap. Quelques chiffres devraient ouvrir les yeux et les esprits (et non, ce n'est pas une fracture du crâne) :

- 1 femme sur 3 dans le monde a déjà été victime de violences en ligne, selon ONU Femmes. 73 % des femmes journalistes ont déjà subi des violences en ligne (UNESCO, 2021). Les adolescentes sont 8 fois plus susceptibles que les garçons d'être victimes de revenge porn (Cyber Civil Rights Initiative).
- 40 % des jeunes LGBTQIA+ déclarent avoir subi du cyberharcèlement (GLSEN, 2023). Les attaques portent souvent sur l'orientation sexuelle, l'identité de genre, les pronoms. Pour les personnes trans, les violences peuvent aller jusqu'au outing forcé, à la moquerie de leur apparence ou à la diffusion de photos "avant/après transition".
- 70 % des ados disent avoir déjà assisté à une situation de cyberharcèlement (UNICEF 2022).
- Les propos racistes, antisémites, islamophobes ou afrophobes abondent sur les réseaux sociaux. Les femmes racisées subissent des violences croisées : sexistes et racistes. D'ailleurs, les femmes noires sont 84 % plus susceptibles d'être mentionnées dans des tweets violents que les femmes blanches (Amnesty, 2020)
- Les personnes en situation de handicap sont peu représentées dans les chiffres, mais très concernées. Les moqueries, les vidéos humiliantes ou le validisme ordinaire (ex : "T'es trop moche pour être harcelée") circulent plus vite qu'on ne le croit.

QUAND L'IA DEVIENT COMPLICE DES VIOLENCES

On aurait pu espérer que l'arrivée de l'intelligence artificielle serve à désamorcer les violences en ligne. Qu'elle aide à les détecter plus vite. À protéger les victimes. À bloquer les contenus problématiques avant qu'ils ne circulent. Hélas, pour l'instant, c'est plutôt l'inverse : elle les aggrave souvent plus qu'elle ne les résout. Ce que l'IA a changé, ce n'est pas juste une évolution technique : c'est un changement de nature.

Je vous le disais plus haut, on a décidé de concentrer ce mag sur le phénomène des deepnudes car clairement, il fait peur et touche à une intimité extrême. Pour celles et ceux qui ne le savent pas, les deepnudes sont des contenus générés par intelligence artificielle, à partir d'une simple photo de la personne ciblée. Une photo ordinaire, publique, issue d'un compte de réseau social, d'un portrait professionnel (hello LinkedIn) ou d'un selfie (option sirotage de Pina Colada) pendant les vacances. Une photo habillée. Sans ambiguïté. Une photo de classe ou une photo avec vos collègues fera bien l'affaire.

Et pourtant, cette image suffit. Injectée dans une IA spécialisée, elle devient la base d'une représentation nue du corps. Un corps imaginaire, mais crédible. Reconstitué selon des critères prédictifs (proportions, texture de peau, formes supposées). Le résultat : une image ou une vidéo pornographique hyperréaliste, qui semble vraie. Une nudité fabriquée, mais profondément invasive. Child Focus indique d'ailleurs qu'un nombre croissant de dossiers impliquant l'IA sont ouverts. Une étude belge confirme également la propagation du phénomène (environ 14% des jeunes ont déjà reçu un deepnude).

C'est là que réside le basculement. Car même en l'absence d'image intime initiale, le corps de la personne est exposé. Pas son corps réel, non. Mais une version plausible, générée à partir de son apparence, d'après ce que l'algorithme "devine". Une nudité qu'elle n'a jamais montrée, mais qui est désormais visible. Partageable. Téléchargeable. Préjudiciable.

DEEPNUDES : COMPRENDRE L'IMPACT PSYCHOLOGIQUE

Quand on parlait de revenge porn il y a quelques années, j'entendais (et j'entends encore) de nombreuses réactions: "Mais elle l'a bien cherché", "Elle n'avait qu'à pas envoyer de photo d'elle nue", "C'était un gros risque mais elle l'a quand même pris", etc etc etc etc. En plus de faire porter la responsabilité de la violence à la victime (slut shaming) et d'accentuer son sentiment de culpabilité, l'IA ici donne une autre dimension au phénomène et réfute finalement ces justifications bidons que certain·es trouvaient aux agresseurs. Le deepnude est entièrement fake et s'inspire d'une simple photo.

Du coup, vous vous doutez bien qu'on entend de nouvelles répliques toutes faites "ouiii mais, c'est pas si grave. C'est pas vraiment son corps. Il ne faut pas qu'elle soit gênée." REALLY ? Okay.

Est-ce que vous ne seriez “pas gêné·e” vous :

- Si votre patron·ne tombe sur une image de vous à poil ? Même si ce n’est pas votre vrai corps, hein. Juste votre tête collée à un torse imaginé par une IA.
- Si vos collègues, vos élèves, vos étudiant·es partagent une vidéo porno où on pense que c’est vous ?
- Si votre ex reçoit ce genre de photo de “vous”, envoyée par un faux compte avec votre prénom ?
- Si vos enfants ou leurs copains/copines tombent sur une image deepnude de “maman” en scrollant sur un forum ?
- Si votre nom apparaît dans un thread Telegram listant les “meufs générées”, triées par ville, par âge, par profession ?
- Si votre photo LinkedIn sert de base à un fake porno qui tourne sur Discord avec des commentaires de type “elle doit aimer ça, la compta hein 😏” ?

Voilà voilà. Empathie first.

Deux expertes nous expliquent pourquoi, même truquées, les images de deepnudes provoquent un choc bien réel chez les victimes : Marine Manard, docteure en sciences psychologiques, neuropsy, auteure et fondatrice du PSST (Parentalité sans tabou) et Marine Ghuys, psychologue spécialisée en périnatalité et fondatrice aussi du compte Instagram “La maternité sans tabous”.

1. UN CHOC TRAUMATIQUE IMMÉDIAT

Se voir mis·e en scène dans une vidéo ou une image sexuelle que l’on n’a jamais tournée peut déclencher un état de choc violent. Contrairement à une image réelle non consentie, ici il n’existe aucun souvenir, aucune expérience préalable. La surprise, l’incompréhension et la sidération sont totales. Ce choc est d’autant plus fort si l’acte représenté est quelque chose que la personne n’a jamais pratiqué, ou qu’elle aurait toujours refusé, car il confronte directement la victime à une violation imaginaire de son intimité.

2. ATTEINTE À L’IDENTITÉ ET DÉPOSSESSION DU CORPS

Un deepfake sexuel ne touche pas uniquement à la réputation publique : il agit aussi sur la manière dont on se perçoit soi-même. Le visage, les expressions, parfois certains détails physiques sont reconnaissables, mais le corps, les gestes ou le contexte ne correspondent pas à la réalité. Cette appropriation visuelle donne la sensation que son corps ne nous appartient plus. L’identité privée (comment je me vois) et l’identité publique (comment les autres me voient) se retrouvent bousculées en même temps, créant une impression de perte totale de contrôle.

3.DISSONANCE, FAUX SOUVENIRS ET CULPABILITÉ

Face à ce type d'image, certaines victimes décrivent un réflexe de vérification incessant, proche d'un "jeu des sept erreurs" : comparer chaque partie du corps visible pour savoir si c'est vraiment la leur. Cette dissonance entre perception intime et image extérieure entretient l'anxiété et la confusion. Les recherches sur les faux souvenirs (Loftus & Pickrell, Wade et al.) montrent que notre cerveau peut intégrer comme réel un événement entièrement inventé s'il est présenté de manière crédible. Même en sachant que le deepfake est faux, un doute ou une culpabilité implicite peuvent s'installer, comme si l'acte représenté avait pu avoir lieu. Cette culpabilité, souvent nourrie par la honte ou le jugement social, fragilise encore l'estime de soi et complique la reconstruction.

En gros, il s'agit d'une effraction dans l'intimité par la simulation. Ce que l'intelligence artificielle rend possible, ce n'est pas seulement l'accélération d'un processus, c'est un changement de paradigme : il n'est plus nécessaire que quelque chose ait eu lieu pour que la personne en subisse les conséquences. Il n'est plus nécessaire d'avoir partagé un contenu pour être punie de l'avoir soi-disant fait.

L'IA crée une fausse réalité, mais la violence, elle, est bien réelle. L'IA n'a pas inventé les violences sexuelles, ni le harcèlement numérique. Cependant, elle leur donne une nouvelle intensité. Elle rend la transgression plus rapide, plus facile, plus anonyme, et plus massive. Elle démultiplie les canaux de diffusion. Elle sème le doute. Elle brouille les repères. Et elle retire aux victimes leurs moyens de se défendre : que répondre, quand ce qui vous salit n'a jamais existé ?

Il est urgent de reconnaître cette nouvelle forme de violence, de la nommer, de la documenter, et surtout, de la considérer avec la gravité qu'elle impose. Le fait que les contenus soient fabriqués ne les rend pas moins destructeurs. Le fait que la nudité soit simulée ne la rend pas moins violente.

MINIMISER EST UN RÉFLEXE QUI FAIT PLUS DE MAL QUE DE BIEN

Face à une victime de deepnude, certains proches ou collègues réagissent en minimisant : *« C'est pas si grave », « Passe à autre chose »*. Ce n'est pas toujours de la mauvaise volonté. Parfois, c'est même pensé comme un geste protecteur. Marine Ghuys dit même qu'à court terme, cette réaction peut donner l'impression d'alléger le choc. L'idée implicite, c'est de réduire l'intensité émotionnelle, d'éviter de "raviver" la blessure. Les proches croient souvent qu'en mettant de côté l'événement, la victime retrouvera plus vite une forme de normalité. Et dans certains cas, sur le moment, cette stratégie peut effectivement apporter un soulagement temporaire (Lepore et al., 1996 ; Major et al., 1990).

En outre, à long terme, l'effet est tout autre. Marine Manard nous raconte que la minimisation empêche la **validation émotionnelle** : la victime n'entend jamais que son ressenti est légitime, ce qui freine la reconstruction. Elle favorise aussi l'**augmentation de la détresse psychologique** : plus d'anxiété, plus de honte, plus de culpabilité, comme le montrent les recherches sur la victimisation secondaire (Campbell & Raja, 1999 ; Orth, 2002). Enfin, elle isole : quand on sent que ses émotions ne sont pas reconnues, on a moins envie de chercher du soutien.

Nos deux expertes proposent une prise en charge émotionnelle et psychologique, et sans minimiser la souffrance.

1. La première étape, c'est de **valider les émotions**. Qu'il s'agisse d'une image réelle ou générée par IA, l'impact est tangible. La peur, la colère, la honte ou l'anxiété sont légitimes et doivent être reconnues.
2. Éviter la minimisation est tout aussi crucial. Les phrases comme *« ce n'est pas ton vrai corps »* ou *« passe à autre chose »*, même prononcées avec l'intention de rassurer, aggravent la souffrance. Elles peuvent entraîner ce que les psychologues appellent une **victimisation secondaire** : la personne doit alors lutter à la fois contre l'agression initiale et contre le déni ou la banalisation de ce qu'elle subit.
3. L'entourage a un rôle clé. En informant les personnes qui entourent la victime et en les sensibilisant, via de la **psychoéducation**, on peut leur donner les bons réflexes de soutien et leur expliquer pourquoi certains propos blessent.
4. Dans la prise en charge, la règle est simple : considérer la situation comme un **traumatisme réel**, et non comme un "faux problème" parce que l'image n'existe pas physiquement. Cela implique d'offrir le même type d'accompagnement psychologique que pour une agression sexuelle filmée sans consentement.
5. Enfin, il est important de travailler sur la **culpabilité**. La victime n'est en rien responsable. Comprendre les mécanismes qui mènent à l'agression permet souvent de déconstruire l'auto-blâme. Quant au suivi, il doit être adapté à chaque personne : thérapies cognitives et comportementales, hypnose, pleine conscience, sophrologie… L'essentiel est d'expérimenter, garder ce qui aide, et écarter ce qui ne fonctionne pas.

DEEPNUDES : QUE FAIRE LÉGALEMENT ?

En Belgique, la loi ne fait pas de différence : qu'elle soit réelle ou générée par IA, une image intime diffusée sans accord est considérée comme du voyeurisme ou de la diffusion non-consentie. C'est une infraction pénale, et les démarches possibles sont les mêmes. Child Focus, avec qui on travaille régulièrement, aide à naviguer dans des démarches officielles. Niels Van Paemel, policy advisor, rappelle les étapes clés :

1. **Rassemblez les preuves**: conservez des captures d'écran, notez les liens, les dates et les heures. Ces éléments sont indispensables pour appuyer toute demande de suppression ou plainte.
2. **Demandez la suppression**: chaque site ou plateforme est censé proposer un système de signalement. Les moteurs de recherche comme Google disposent aussi de formulaires spécifiques pour faire disparaître les contenus explicites des résultats.
3. **Signalez** sur les réseaux sociaux via les formulaires internes.
4. **Portez plainte** à la police : on notera tout de même que c'est possible mais que ce n'est pas toujours simple : la police n'est pas encore formée partout à ce type de situation, ce qui peut rendre l'accueil difficile. Child Focus travaille à améliorer cette prise en charge et peut aider à préparer la plainte, pour que la victime ne se retrouve pas seule face à l'incompréhension ou au scepticisme.
5. **Contactez Child Focus** (116000) pour un accompagnement juridique et psychosocial (mineur·es et proches).
6. **Protégez ses comptes** : enfin, il convient de bloquer les comptes harceleurs et de renforcer ses propres paramètres de confidentialité. Cela ne supprime pas l'image, mais limite l'exposition et protège contre de nouvelles attaques.

DU COUP, L'IA PEUT-ELLE NOUS PROTÉGER ?

On y arrive ! Et c'est quand même une sacrée question. Depuis le début de ce numéro, on évoque surtout en quoi l'IA rajoute une couche problématique aux situations déjà préoccupantes du cyber-harcèlement. On pourrait alors se demander si elle pourrait servir à l'autre camp : celui des victimes, des associations, des modérateur·ices, des éducateur·ices, des plateformes (quand elles veulent bien s'en donner la peine). Mais ça demande un gros "si".

Ce que l'IA peut déjà (parfois faire) :

- **Détecter automatiquement certains contenus violents** : insultes, menaces, photos explicites... Des IA sont entraînées à reconnaître des schémas de harcèlement ou de langage sexiste (mais elles ont encore du mal avec le second degré, les emojis, les détournements ou l'humour raciste déguisé. On en parle un peu après).

- **Filtrer les deepfakes/deepnudes** : des outils comme Deepware, Sensity ou Hive AI peuvent analyser une vidéo ou une image pour détecter des traces de génération artificielle. Certains sont intégrés à des outils journalistiques ou policiers.

- **Accompagner les victimes dans la prise de parole** : des chatbots d'accompagnement existent, comme Wysa ou Woebot (dans d'autres contextes, souvent en santé mentale), et pourraient être adaptés pour accueillir la parole de victimes et leur proposer des ressources fiables.

- G**énérer des contenus de prévention** : car oui, paradoxalement, on peut utiliser l'IA pour créer des campagnes éducatives, générer des scripts de sensibilisation, ou même former les modérateur·ices aux nouveaux types de discours toxiques.

MAIS

Il y a cette idée, presque rassurante, que l'intelligence artificielle finira par nous soulager. Qu'elle triera les propos violents à notre place, floutera les images problématiques, bloquera les comptes toxiques avant même qu'ils n'agissent. Qu'elle saura, à force d'apprentissage, faire le tri entre l'ironie et la haine, entre l'humour noir et la violence pure. En réalité, cette promesse reste largement théorique.

Derrière les grandes déclarations et les interfaces bienveillantes "Nous avons mis en place une modératon par IA", il y a des intelligences artificielles qui ne comprennent pas ce qu'elles lisent, qui suppriment au hasard, qui censurent le mot "trans" ou "féministe" et laissent passer des menaces voilées ou des insultes codées. Il y a des plateformes comme Twitch qui affirment avoir des systèmes de filtrage "ultra performants", alors même que les raids haineux contre les streamers·euses racisé·es ou queer continuent chaque semaine. Il y a des IA qui censurent un message militant parce qu'il contient le mot "viol", mais qui ne réagissent pas à un thread Telegram d'agresseurs organisés.

Après déjà 3 numéros, vous le savez : l'IA ne pense pas, elle calcule. Elle détecte des formes, des fréquences, des anomalies, des mots interdits. Mais elle ne sait pas lire entre les lignes. Elle ne sent pas l'ironie, le contexte, la répétition douce mais violente d'un harcèlement insidieux. Elle ne voit pas que cette "private joke" est une attaque. Que ce message privé est un piège. Que cette vidéo est une menace. Elle applique un protocole. Et dans ce protocole, il y a ce que des développeurs ont décidé d'encoder, ce qui est un problème en soi, car leurs références, leur langue, leurs biais, leur culture façonnent l'outil. Et parfois, l'outil devient une arme au service de ce qu'il prétend combattre. On vous invite à écouter l'épisode 4 d'OK Mila, avec Chloé Tran Phu (Qui c'est qu'a fait la comm? héhé).

Aujourd'hui, des plateformes entières (Twitch, X/Twitter, Insta, même YouTube) se cachent derrière l'IA. Elles ont "mis en place un filtre", "déployé une couche de modération automatique", "optimisé leur machine d'analyse". Et donc, si le harcèlement passe, si les images non consenties circulent, si les personnes ciblées ne sont pas protégées, ce n'est pas leur faute. C'est "la technologie qui n'est pas encore parfaite". Comme si elles n'étaient pas responsables de ce qu'elles hébergent. Comme si la machine les absout. Ce qu'on oublie dans cette histoire, c'est que la violence, elle, reste humaine. Et que c'est justement parce qu'elle est humaine qu'elle prend parfois des formes tordues, indirectes, sournoises, et que l'IA ne sait pas les reconnaître. Ce qu'il faudrait, ce n'est pas juste une IA plus puissante. C'est une IA encadrée, assistée, humanisée et surtout, un cadre politique et social qui ne laisse pas la machine décider seule de ce qui est acceptable ou pas. Il faudrait des humains formés, présents, engagés. Des collectifs d'usager·ères écouté·es. Des plateformes qui assument leur rôle de modération, non pas en le déléguant entièrement, mais en l'assumant comme une responsabilité éthique.

FOCUS INSPIRATION

ZOOM SUR DATAFORETHIC / NETETHIC

Développée en France, DataForEthic propose un système de détection automatisée des cyberviolences dans les environnements scolaires. Sa plateforme Netethic Éducation agit comme une forme de surveillance éthique, capable d'analyser les messages, les échanges et les images partagés via les ENT (espaces numériques de travail), les messageries internes ou les plateformes collaboratives entre élèves. L'outil analyse les contenus en temps réel et déclenche des alertes à destination de l'équipe éducative en cas de propos sexistes, homophobes, menaçants, dégradants ou violents.

Pourquoi c'est intéressant ? Parce que c'est l'une des seules solutions IA concrètes pensées non pas pour remplacer la modération humaine, mais pour la soutenir intelligemment. Elle combine :

- une veille algorithmique sur plus de 30 types de violences verbales ou symboliques
- un système de signalement anonyme à destination des élèves
- une logique d'accompagnement éducatif, avec des retours humains et plans d'action adaptés.

Évidemment, ce concept est adaptable à d'autres milieux. Le plus gros intérêt de Netethic, c'est que sa logique n'est pas exclusivement scolaire. Ce type d'outil pourrait être répliqué ou inspirer des alternatives IA dans d'autres espaces en ligne, aujourd'hui saturés de violences :

- Streaming et jeux vidéo : imaginons un équivalent de Netethic intégré à Twitch ou à Discord. Une IA ne modérant pas uniquement par mots interdits, mais détectant des dynamiques de raids haineux (pics d'entrées sur un stream, vagues de commentaires coordonnés), des harcèlements ciblés, ou des private jokes répétées à caractère raciste/sexiste.
- Réseaux professionnels : dans les environnements de travail (Slack, Teams...), une IA pourrait signaler les propos déplacés, le harcèlement passif-agressif, les remarques sexistes ou validistes. Couplée à une cellule RH ou égalité, l'IA devient alors un déclencheur d'écoute plutôt qu'un outil de surveillance.
- Réseaux sociaux grand public : plutôt que de laisser Meta, X ou TikTok prétendre modérer avec une IA "magique" sans transparence, on pourrait imaginer des modèles inspirés de Netethic, où la détection algorithmique s'accompagne d'un ancrage dans les vécus des victimes et d'une vraie redevabilité. Des collectifs d'usager·ères pourraient participer à l'élaboration des critères, au contrôle des biais, à l'analyse des dérives. Mais attention dans ce cas grand public : cette implication ne doit en aucun cas servir de prétexte à désengager les plateformes. La responsabilité légale, politique et structurelle doit rester entre les mains de celles qui hébergent, diffusent, et monétisent (genre, on ban définitivement des harceleurs). L'outil IA, même collaboratif, ne doit pas devenir un paravent. Ce serait une double peine : être harcelé·e, puis devoir modérer soi-même la haine qu'on reçoit.

ZOOM SUR CHILD FOCUS & CO

Comme on le disait un peu avant, Child Focus traite des questions de sexting non consensuel dont les deepnudes font partie. Lors de notre interview, Niels nous a recommandé plusieurs outils.

116000 : le numéro de Child Focus

La ligne d'assistance de Child Focus est ouverte 24h/24 et 7j/7. C'est là qu'une victime (ou un parent) peut :

- obtenir un soutien émotionnel et psychologique immédiat,
- recevoir des conseils concrets pour constituer un dossier (quoi collecter, comment signaler),
- être orienté·e vers la police ou un service juridique.

Arachnid

Arachnid est une base internationale d'empreintes numériques d'images illégales, alimentée par des ONG et Interpol. Child Focus y envoie les contenus signalés en Belgique, ce qui permet leur détection et suppression, même s'ils réapparaissent sur d'autres sites ou serveurs à l'étranger.

Take It Down / StopNCII.org

Imaginez que votre photo intime (réelle ou générée par IA) soit une pièce unique, avec une empreinte digitale que personne d'autre n'a. Take It Down (pour les mineur·es) et StopNCII.org (pour les adultes) permettent de créer cette "empreinte"qu'on appelle un hash, directement depuis votre téléphone ou votre ordi... sans jamais envoyer la photo. Une fois l'empreinte créée, elle est transmise aux grosses plateformes partenaires (Meta, TikTok, OnlyFans, Reddit, Yubo...). Si quelqu'un tente de publier l'image (ou une version retouchée), la plateforme compare les empreintes : si ça matche, blocage immédiat.

Suppression ciblée sur les plateformes et Google

Child Focus conseille et assiste pour utiliser les formulaires internes aux réseaux sociaux et déclencher une suppression dans Google Images / résultats de recherche.

Faire full prévention et intégrer l'éducation aux médias dans les programmes scolaires

Bien sûr, aucun outil ne remplace une vraie politique de prévention, d'éducation et d'écoute. L'éducation aux médias doit commencer tôt, dès l'école primaire, pour apprendre la bienveillance en ligne et le respect de soi et des autres. Au fil des années (si l'EAM était intégrée réellement aux programmes scolaires), ces bases permettent d'aborder des sujets plus complexes : cyber-harcèlement, deepfakes, diffusion non consentie d'images, violences numériques... et donner à chacun·e les bons réflexes avant qu'un problème n'arrive.

Child Focus propose "Click Safe", une formation sur un ou deux jours à destination des enseignant·es, des éducateur·rices, de la police, des PMS, etc. Avec cette formation pédagogique, Child Focus souhaite aider les professionnel·les à aborder le sujet du bon usage d'internet avec des enfants et des adolescent·es. La formation fournit des informations générales sur la façon dont les jeunes gèrent internet et les risques qui y sont associés, mais propose également de nombreux supports pédagogiques prêts à l'emploi. Une réflexion est également abordée sur la mise en place d'une politique scolaire en la matière. Une véritable boîte à outils pédagogiques.

Action Médias Jeunes propose des ateliers, animations et formations en éducation aux médias, dont le cyher-harcèlement fait partie : www.actionmediasjeunes.be

Infor Jeunes propose des animations comme **« Je réfléchis puis je clique »** pour prévenir le harcèlement et le cyber-harcèlement. En plus, la cellule de Bruxelles propose un soutien direct aux jeunes confrontés au harcèlement : www.inforjeunes.be

Le **CRIH (Centre de Référence et d'Intervention Harcèlement),** présent dans une trentaine de communes en Hainaut accompagne enfants, parents et professionnel·les face au harcèlement scolaire (y compris en ligne) : à retrouver sur leur page Facebook.

Le **Centre ReSIS** est une association proposant formations et conférences pour prévenir, détecter et agir contre le harcèlement et cyberharcèlement en milieu scolaire, via la Méthode de la Préoccupation Partagée : www.centreresis.org/

Bibliothèques Sans Frontières et son kit pédagogique « Les Cyber Héros » sensibilisent les enfants de 8 à 13 ans à la sécurité numérique, à la citoyenneté en ligne et au cyberharcèlement, de façon ludique. Ils proposent aussi des formations pour enseignant·es, ainsi que des ateliers « Cyber School » en classe : www.bibliosansfrontieres.be/

L'**Université de Paix** dispense des formations et animations pour apprendre à gérer les conflits, améliorer la communication et renforcer la coopération, dès la maternelle jusqu'aux adultes. Leur approche inclut la prévention et la gestion du harcèlement, y compris en ligne, en travaillant sur l'estime de soi, l'empathie et les compétences relationnelles : www.universitedepaix.org/

REMERCIEMENTS

La team OK Mila, Gaëlle Defeyt, Anne-Sophie Skit, Julie Mouvet, François d'Alcamo, Marius Adam, Francisco Luzemo, Elisabeth Bois d'Enghien, Cassi Henaff

Le CSEM, le Conseil Supérieur de l'Education aux Médias sans qui le projet OK Mila, n'aurait pas pu voir le jour.

Tous nos contributeurs et contributrices de feu :

- Chloé Tran Phu de Media Animation et du Collectif Witch Gamez https://witchgamez.com/
- Florence Hainaut https://www.instagram.com/florencehainaut/
- Marine Manard (Parentalité sans tabou) : https://www.psst-magazine.be/
- Marine Ghuys (La maternité sans tabous) : https://www.instagram.com/la_maternite_sans_tabous/
- Niels Van Paemal de Child Focus https://childfocus.be/fr-be/
- Rashka Barbare https://www.instagram.com/game_girl.podcast/

Ce projet a été créé par Tapage Studio www.ok-mila.be - hello@tapage.studio - www.tapage.studio

BIBLIOGRAPHIE ET CITOGRAPHIE

Loftus, E. F. (2003). *Make-Believe Memories*. *American Psychologist*, 58(11), 867–873 **:** Étude fondatrice sur la création de faux souvenirs, montrant que des images ou récits inventés peuvent être intégrés à la mémoire comme s'ils étaient réels.

Shaw, J. (2020). *Do False Memories Look Real? Evidence That People Struggle to Identify Rich False Memories of Committing Crime and Other Emotional Events*. *Frontiers in Psychology* **:** Montre que les faux souvenirs émotionnels sont souvent indiscernables pour les personnes concernées, renforçant la difficulté à distinguer le vrai du faux.

Lepore, S. J., Silver, R. C., Wortman, C. B., & Wayment, H. A. (1996). *Social constraints, intrusive thoughts, and depressive symptoms among bereaved mothers* : Montre que les contraintes sociales à parler de son vécu (dont la minimisation) peuvent réduire l'expression émotionnelle et augmenter les symptômes dépressifs.

Major, B., Richards, C., Cooper, M. L., Cozzarelli, C., & Zubek, J. (1990). *Personal resilience, cognitive appraisals, and coping: An integrative model of adjustment to abortion* : Explique comment certaines stratégies d'adaptation, comme la minimisation, peuvent atténuer la détresse immédiate mais nuire à la résolution à long terme.

Linehan, M. M. (1993). *Cognitive-Behavioral Treatment of Borderline Personality Disorder.* : Conceptualise l'importance de la validation émotionnelle pour traiter les traumatismes et éviter la détresse prolongée.

Campbell, R., & Raja, S. (1999). *Secondary victimization of rape victims: Insights from mental health professionals who treat survivors of violence* : Montre que la victimisation secondaire (dont la minimisation) est associée à une augmentation de l'anxiété, de la honte et de la culpabilité.

Ullman, S. E. (2010). *Talking about Sexual Assault: Society's Response to Survivors* **:** Souligne que le manque de reconnaissance et la minimisation sociale peuvent décourager la recherche de soutien.

Bates, S. (2017). *Revenge porn and mental health: A qualitative analysis of the mental health effects of revenge porn on female survivors* **:** Montre que la banalisation et la minimisation publiques des violences sexuelles numériques renforcent l'injustice perçue et la détresse psychologique.

ONU Femmes. (2024). *Violence against women* https://www.unwomen.org/fr/what-we-do/ending-violence-against-women/facts-and-figures

UNESCO & ICFJ. (2021). *The Chilling: Global trends in online violence against women journalists* https://unesdoc.unesco.org/ark:/48223/pf0000377224

Cyber Civil Rights Initiative. (2023). *Image-Based Sexual Abuse: Statistics* https://www.cybercivilrights.org

GLSEN. (2023). *2023 National School Climate Survey* https://www.glsen.org/research/school-climate-survey

UNICEF. (2019). *Are children safe in the digital world? Global Kids Online findings* https://www.unicef.org/end-violence/how-to-stop-cyberbullying

Amnesty International. (2020). *Toxic Twitter – Women of Colour* https://www.amnesty.org/en/latest/research/2020/03/toxic-twitter-2020/

France Bleu. *"Boomer traps" : comment les arnaqueurs approchent les personnes âgées grâce aux fausses images* https://www.francebleu.fr/infos/societe/boomer-traps-comment-les-arnaqueurs-approchent-les-personnes-agees-grace-aux-fausses-images-5511780

RTBF. *Arnaque aux deepfakes et aux faux articles : que fait Facebook contre les publicités mensongères utilisant l'image de personnalités?* https://www.rtbf.be/article/arnaque-aux-deepfakes-et-aux-faux-articles-que-fait-facebook-contre-les-publicites-mensongeres-utilisant-l-image-de-personnalites-11569572

MACSF. *Intelligence artificielle : escroqueries, nouvelles menaces* https://www.macsf.fr/actualites/intelligence-artificielle-escroqueries-nouvelles-menaces

Police fédérale belge. *Attention à l'arnaque aux fausses factures : jusqu'à 130 000 euros dérobés* https://www.police.be/5998/fr/actualites/attention-a-larnaque-aux-fausses-factures-jusqua-130-000-euros-derobes

Ludomag. *DataforEthic : une solution anti-harcèlement basée sur l'IA* https://www.ludomag.com/2024/07/03/dataforethic-une-solution-anti-harcelement-basee-sur-lia/

Child Focus. *Informations de prévention sur les deepnudes* https://childfocus.be/fr-be/Exploitation-Sexuelle/Deepnudes

RTBF. *L artificielle pour déshabiller des femmes : c'est illégal et vraiment inquiétant, y compris en Belgique* https://www.rtbf.be/article/l-intelligence-artificielle-pour-deshabiller-des-femmes-c-est-illegal-et-vraiment-inquietant-y-compris-en-belgique-11259804

I**nstitut pour l'égalité des femmes et des hommes.** *Étude sur la diffusion des deepnudes parmi les jeunes Belges* : https://igvm-iefh.belgium.be/sites/default/files/les_deepnudes_parmi_les_jeunes_belges.pdf

ViolencesSexuellesEnLigne.be. *Deepfakes sexuels : situation-type et conseils* https://www.violencessexuellesenligne.be/situation-10-complet-deepfake/

I**nstitut pour l'égalité des femmes et des hommes.** *Page thématique "Deepnudes"* https://igvm-iefh.belgium.be/fr/themes/violences-sexuelles-numeriques/deepnudes

LN24. *Les deepnudes : les nouveaux ravages de l'IA* https://www.ln24.be/videos/2025/02/13/les-deepnudes-les-nouveaux-ravages-de-l-ia-xm0frfq/

OK MILA

L'intelligence artificielle bouleverse notre manière de nous informer, de créer et d'interagir. Avec OK Mila, nous décryptons ses impacts éthiques et sociétaux à travers six thématiques clés :

- **IA et cyber-harcèlement (septembre 2025)** : Le cyber-harcèlement se renouvelle sans cesse à travers de nouvelles formes : deepfakes pornographiques, création de nudes, faux chatbots d'harcèlement, modération douteuse sur certaines plateformes comme Twitch. Comment sensibiliser pour mieux prévenir ces violences et, surtout, quelles solutions proposer aux victimes pour réagir efficacement ?
- **IA et politique (octobre 2025) :** La propagande politique ne concerne pas uniquement les régimes autoritaires. Elle s'exerce aussi chez nous, à travers les algorithmes de big data, les chatbots des partis ou encore l'ajustement permanent des stratégies de communication. Comment l'intelligence artificielle impacte-t-elle concrètement nos démocraties ?
- **IA et créativité (novembre 2025)** : Si l'IA suscite des inquiétudes légitimes du côté des artistes, elle ouvre aussi des perspectives inédites en matière de création. Elle soulève en parallèle des questions essentielles sur la protection des œuvres et les droits d'auteur·ices. Comment utiliser l'IA pour stimuler sa créativité tout en respectant celles et ceux qui créent ?

Les 3 premières thématiques sont sorties en mars, avril et juin et abordaient : l'IA et la désinformation, l'IA et la diversité et l'IA et l'écologie.

Un projet multimodal pour toucher un large public

OK Mila, c'est trois formats interconnectés pour explorer ces enjeux en profondeur et développer un esprit critique face à l'IA :

- Un podcast (12 épisodes) : une thématique et 2 interventions, un sujet et deux volets car les discussions autour de l'IA et de l'éthique révèlent de nombreuses questions.
- Des vidéos courtes : des capsules percutantes et accessibles pour sensibiliser un large public autour d'informations concrètes.
- Un magazine papier (6 numéros) : un format A2 pliable avec un lexique IA, des interviews et des QR codes vers des ressources supplémentaires, diffusé dans les écoles, maisons de jeunes, AMO et centres culturels.

Où nous retrouver ? ▶ Écoutez nos podcasts sur **Spotify, Apple Podcasts, Deezer et toutes les plateformes d'écoute**. ▶ Retrouvez nos vidéos et ressources sur **www.ok-mila-eam.be**. ▶ Suivez-nous sur **Instagram** pour ne rien manquer !

OK MILA MAG

UN PROJET D'ÉDUCATION AUX MÉDIAS QUI QUESTIONNE L'ÉTHIQUE ET L'INTELLIGENCE ARTIFICIELLE

N°04 · IA ET CYBER-HARCÈLEMENT