Les errements sexistes de **l'intelligence artificielle**

Une proportion effrayante de deepfakes, contenus créés grâce à l'intelligence artificielle, sont pornographiques et non consentis. Pour réguler le partage de ces images, les plateformes et les autorités françaises sont à la traîne. ROMAIN HAILLARD

Juste deux mots-clés dans un moteur de recherche, et un site dédié au partage de contenus pornographiques générés par intelligence artificielle (IA) nous ouvre ses portes. Sur ce forum, un espace de discussion affiche la possibilité de générer des images à partir de personnalités francophones. Son créateur indique réaliser des « fakes ». Habitué à opérer laborieusement avec de simples logiciels de montage, *« uniquement dans le but de satisfaire des fantasmes visuels »*, il découvre les joies du progrès technologique. Soit il remplace le visage d'une actrice d'une vidéo pornographique par le visage d'une autre femme, le face-swapping, soit il laisse une IA remplacer le corps habillé d'une femme par un corps nu. Les contenus sont synthétiques mais vraisemblables. L'artisan demande aux membres du forum de simplement lui fournir le nom d'une personnalité, des photos de celle-ci et la situation dans laquelle elle doit se retrouver. Un internaute lui propose une Youtubeuse, demande de *« juste la déshabiller »* et indique qu'il la trouve *« très souriante »* en précisant : *« Je l'adore. »*

Dans ce rapide échange entre hommes, le naturel de la situation sidère. Une personne propose un service à une autre et prend même quelques précautions morales. Il n'est pas question d'utiliser ses services pour se venger d'une femme, de l'utiliser pour des personnes inconnues, encore moins des mineures. Il ne s'agit pas non plus de *« nuire à quelqu'un »*, mais d'un usage purement personnel. Rien n'est grave. Un autre site compile plus de dix mille images de femmes publiques françaises : ministres, athlètes, actrices, chanteuses, influenceuses. Un expert anonyme a compilé des chiffres qu'il a confiés au magazine Wired. Ils indiquent une phase d'accélération de la production de contenus de deepfakes à caractère pornographique : *« À la fin de cette année* [2023]*, plus de vidéos auront été produites en un an que toutes les années combinées. »*

/ 99 % des deepfakes concernent des femmes

Le phénomène a commencé en 2017. Samantha Cole, journaliste à Vice puis à 404 Media, l'annonçait : *« Le porno assisté par intelligence artificielle est là et nous sommes tous foutus (1). »* On pourrait dire plutôt « toutes foutues ». Selon une étude de 2023, 98 % de ces contenus sont pornographiques et 99 % concernent des femmes (2). La journaliste évoquait une vidéo où le visage de l'actrice Gal Gadot – elle a joué Wonder Woman – avait remplacé celui d'une actrice X. Déjà, en 2017, ce face-swap était le fruit d'une personne seule, certes avec des compétences, mais à l'aide d'un logiciel libre d'accès. *« L'alerte avait été donnée, mais personne n'a voulu réagir »*, commente Mathilde Saliou, journaliste à Next, site de référence sur les nouvelles technologies.

Il y a un an, Mathilde Saliou signe *Technoféminisme. Comment le numérique aggrave les inégalités* (Grasset). *« Les outils que nous utilisons sont programmés à 84 % par des hommes »*, avance-t-elle, avant de souligner : *« Ils ont des angles morts. Ça ne serait pas problématique s'ils ne pensaient pas que ce qu'ils font est universel. S'il y avait plus de femmes, moins de CSP+ et plus de diversité, alors ces personnes auraient poussé à une prise en compte rapide des violences en ligne. »* Surtout, la modération n'est pas du goût du seul réseau où il est possible de poster du contenu pornographique : X, anciennement Twitter. Mais, depuis l'adoption du règlement européen sur les services numériques, le DSA, les plateformes sont contraintes de livrer quelques chiffres sur leur modération. Elon Musk, monarque absolutiste de

la liberté d'expression, a donc embauché 52 personnes pour s'assurer du respect de la loi et des règles de X pour les publications francophones.

Son réseau ne s'en est pas moins fait rattraper récemment par l'actualité. Plusieurs deepfakes de la chanteuse états-unienne Taylor Swift ont provoqué un emballement médiatique fin janvier. L'une des images a été vue plus de 45 millions de fois avant d'être supprimée. Dépassée par la situation, l'entreprise a décidé de bloquer temporairement la recherche « deepfake Taylor Swift » pour siffler la fin de la partie. D'autres plateformes ont pris la même décision de manière plus définitive. En 2018, Reedit a banni la recherche de deepfakes de son réseau. Les principales plateformes de visionnage de contenus pornographiques ont également bloqué la recherche du mot-clé « deepfake ».

« Les responsabilités appartiennent à ceux qui ont le pouvoir : les plateformes », défend Laure Salmona, cofondatrice de l'association Féministes contre le cyberharcèlement et coautrice avec Ketsia Mutombo du livre *Politiser les cyberviolences. Une lecture intersectionnelle des inégalités de genre sur Internet* (Le Cavalier bleu, 2023). *« Nous arrivons au point où il faudrait ne plus mettre de photos de soi sur les réseaux sociaux*, s'exaspère-t-elle. *Même mettre des vêtements à vendre sur Vinted expose à une mise à nu et à un partage sur des sites pornographiques. »* Une enquête d'Aurore Gayte sur le média en ligne Numerama avait montré l'existence d'un site où étaient compilées des images voyeuristes collectées sur Vinted. Un onglet « Undress AI » redirigeait vers un site qui proposait de passer n'importe quelle image à la moulinette dopée à l'IA pour synthétiser une image sans vêtements.

/ Données poubelles

Avec une pointe de fatalisme, Laure Salmona voit une ouverture dans ce phénomène : *« Cette technologie remet en question la culpabilisation des victimes et des injonctions à maîtriser leurs images intimes. »* Ces images peuvent même parfois être générées automatiquement. Melissa Heikkilä, journaliste à la *MIT Technology Review*, en a été la victime après avoir utilisé Lensa AI. L'application permet de créer, grâce à l'IA, des portraits à partir de selfies. Sur cent images produites à partir de ses photos, trente la représentaient seins nus, dans des tenues courtes ou dans des positions sexualisées. Selon elle, après comparaison avec une de ses collègues, blanche, ce sont ses origines asiatiques qui auraient conduit la machine à adopter ce biais.

« Sur quelles données reposent ces modèles ? », réagit Mathilde Saliou. Dans son ouvrage, la journaliste relève cet adage de l'informatique : *« Garbage in, garbage out »*, qui signifie que, si les données sont défectueuses à l'entrée, il y a de fortes chances que la machine ressorte des données poubelles (*garbage*). Lors des négociations autour de l'AI Act, le projet de règlement européen sur l'intelligence artificielle, des associations féministes comme Stop Fisha ont appelé à une plus grande transparence dans les usages des algorithmes et leurs données d'entraînement.

Du côté des autorités françaises, la prise en compte du problème semble lente au démarrage. *« Le gouvernement n'est pas hostile, c'est juste un impensé »*, regrette Jean-Christophe Le Toquin, président de Point de contact, une association spécialisée dans le signalement et le retrait de contenus en ligne. Actuellement, la loi Sren, visant à réguler et à sécuriser l'espace numérique, prévoit plusieurs articles pour mieux prendre en compte dans la loi ces contenus non consentis. Le gouvernement a proposé un amendement au Sénat pour créer un délit de publication de deepfake à caractère sexuel, puni de deux ans d'emprisonnement et de 60 000 euros d'amende. Une étape franchie pour Rachel Flore-Pardo, avocate et fondatrice du collectif StopFisha, qui se bat contre le cybersexisme et les cyberviolences.

/ Retrait automatique de contenus

Mais des problèmes persistent en matière de prise en compte de la parole des victimes. *« Pharos* [portail de signalement des contenus illégaux en ligne] *doit revoir son formulaire : il ne prend pas en compte la diversité des atteintes cybersexistes, notamment le deepfake pornographique. La plateforme doit s'adresser à toutes les victimes »*, oppose l'avocate. Sur onze entrées possibles du formulaire, parmi lesquelles figurent « terrorisme », « acte de cruauté envers les animaux » et « spam », aucune ne fait explicitement mention d'une atteinte sexiste. La fondatrice de Stop Fisha soupire : *« Dans le meilleur des cas, il y a une audience, une condamnation, mais même dans ce cas-là il n'y a pas les outils nécessaires pour s'assurer que le contenu ne sera pas republié. »* Aujourd'hui, les victimes doivent faire ce travail manuellement, contenu par contenu : un travail de Sisyphe.

> « Si tu pouvais juste la déshabiller… » demande un internaute à l'IA.

« Nous pensons à mettre les gens au trou, mais que faisons-nous pour que les victimes soient soutenues ? », se demande Jean-Christophe Le Toquin. Des solutions techniques existent. Un contenu vidéo, une fois passé dans un algorithme dit « de hachage », peut être transformé en une signature unique. L'utilité la plus connue de ce type de fonction est de vérifier l'authenticité d'un fichier avant de le télécharger. S'il a été modifié, la signature change. Dans ce cas précis, il permettrait de vérifier la présence d'un fichier identique dans un grand volume de données, comme un site de partage de vidéos pornographiques. *« L'idée, à terme, c'est de créer une base de "hash", de signatures, et de la mettre à disposition des plateformes pour permettre un retrait en continu »*, imagine le président de Point de contact. Une grande plateforme serait déjà partante pour adopter cette méthode.

« J'espère que dans dix ans, ou même avant, nous nous dirons : mais comment les gens vivaient dans cet environnement numérique ? J'espère que ça sera incompréhensible. » Une semaine après notre entretien avec le président de Point de contact, l'association a vu ses financements interrompus par le gouvernement. Considérée comme l'un des premiers signaleurs de confiance auprès de la plateforme Pharos, elle est amenée à disparaître faute de moyens, et avec elle son projet de retrait automatisé.

Le salut ne viendra pas non plus de l'Union européenne. Le règlement sur l'IA classe les deepfakes comme des systèmes à « risques limités ». Ils feront donc l'objet d'une régulation bien moins lourde que la reconnaissance faciale, par exemple. ●

(1) « AI-Assisted fake porn is here and we're all fucked », Vice, 11 décembre 2017.

(2) « States of deepfakes. Home Security Heroes », étude réalisée sur 95 820 vidéos de deepfake publiées sur plus de 100 sites.