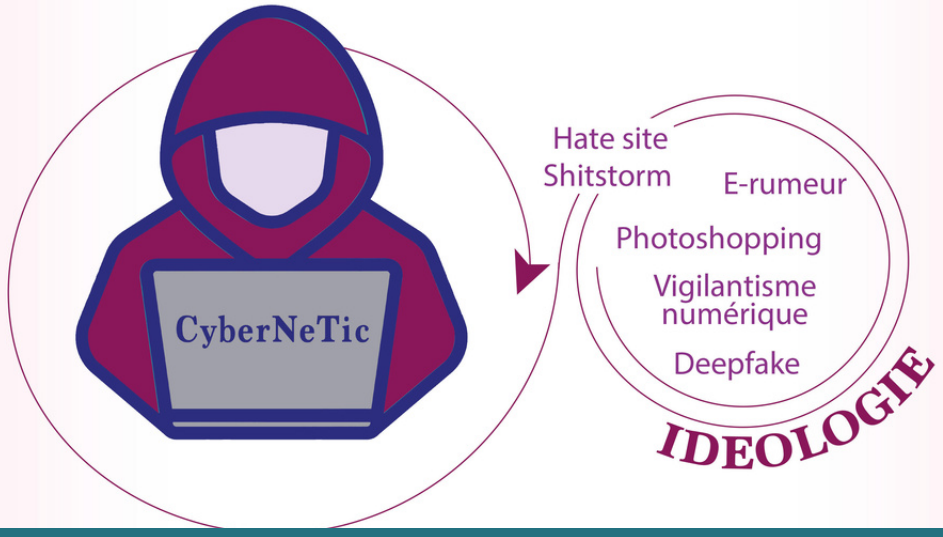
PROJET CYBERNETIC

DEEPFAKE

Etiologie des pratiques de cyberharcèlement



SYNONYMES

- Trucage numérique
- Hypertrucage
- Permutation intelligente de visages et du son
- Infox vidéo
- Vidéotox

Définition

Concept-clé :

Le *deepfake* est une technique de manipulation audiovisuelle qui repose sur l'intelligence artificielle et qui permet d'incruster des visages, d'émuler des voix et des discours ou des gestes dans des vidéos déjà existantes. Souvent utilisé dans les pratiques de cyberharcèlement pour produire des vidéos à caractère pornographique, ce trucage numérique consiste à nuire en détournant l'image d'une personne afin de lui prêter des comportements ou des propos qu'elle n'a pas tenus ou qu'elle ne partage pas.

Le deepfake repose sur une technique de *Machine Learning* qui, à partir d'images déjà fournies, consiste à mettre en compétition deux algorithmes d'apprentissage (Generative Adversarial Network - GAN - soit réseau antagoniste génératif). Le premier algorithme identifié comme "**générateur**" va chercher à créer des contrefaçons les plus crédibles possibles. Le second dit "**discriminateur**" s'applique à détecter les données générées artificiellement le plus efficacement possible. Au fil du temps, les deux algorithmes se perfectionnent dans leur relation d'amélioration continue pour optimiser le niveau de réalisme des images. À un moment donné, le premier algorithme arrive à produire de fausses images sans que le second ne puisse détecter la supercherie. Peuvent se distinguer aujourd'hui plusieurs techniques de synthèse et de montage pour créer un deepfake : le ***faceswap*** (échange de visage), le ***lipsync*** (synchronisation des lèvres), le ***puppeteering*** (expressions faciales et corporelles), etc.

Ce qu'il faut retenir...

En nous intéressant aux facteurs qui contribuent à la propagation des *deepfakes*, 3 composants typiques peuvent être identifiées :

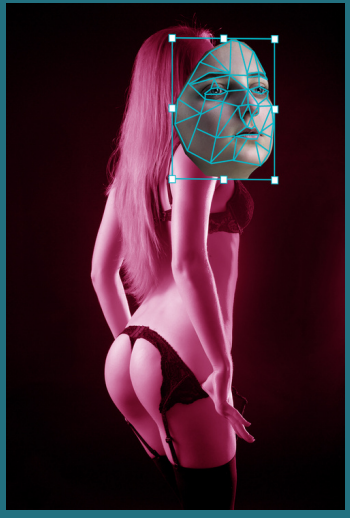
-**L'hypervisibilité** : si les logiciels de trucage numérique sont aujourd'hui faciles d'utilisation, la réalisation d'une vidéotox doit cependant pouvoir s'appuyer sur un nombre important d'**images et de vidéos en ligne** pour se rapprocher le plus possible d'un certain **réalisme**. Cette injonction de visibilité ne peut se dissocier de celle de **modernité** (le fait d'être en phase avec l'actualité, d'être à la mode, de correspondre à une tendance récente), et contribue considérablement à sa propagation sur la toile. Ainsi, plus la personne est **populaire** et **référencée** sur internet, plus elle est susceptible de faire l'objet d'un *deepfake*.

-**L'exclusivité** : plusieurs études et rapports (institut Reuters, european journalism observatory) montrent que les sources officielles suscitent aujourd'hui une certaine **défiance** de la part du grand public (crédibilité des journalistes remise en question, moyens d'information traditionnels boudés par les jeunes générations, etc.). Sont au contraire privilégiés sur internet des sources **alternatives** (Ducol ; 1997) qui se focalisent sur la **révélation**, **l'exclusivité**, le **scoop** (comme la source d'un témoin de l'affaire, un proche du dossier, etc.).

-**L'émotion** : un lien fondamental est établi entre **l'émotion mémorisable** et le **pouvoir de l'image** (Barre ; 2020). Ainsi une vidéo à caractère sexuel, perçue dans sa dimension **transgressive** du tabou, est un élément porteur de **sensations** pour le spectateur. **L'évocation de l'interdit** procure des **émotions contradictoires** pour celui qui regarde, entre l'offense faite à son système de valeurs et la réjouissance par procuration à franchir les limites de cet intime. "*L'information ne cherche pas ici un savoir ni même un voir, mais un faire-voir susceptible de produire directement un croire, indispensable à l'émotion*" (Têtu ; 2004).

> ***Vous vous retrouvez, malgré vous, au cœur d'un film pornographique. On sait que ce n'est pas son corps, mais au bout de 30 à 40 secondes, on en a pourtant l'impression!***

Un exemple concret :



Aux origines...

Une première interprétation du mot ***deepfake*** peut être proposée dans la contraction de l'anglais "deep learning", le système d'apprentissage qui utilise l'intelligence artificielle, et de "fake" qui signifie contrefait. Ainsi pourrions-nous traduire ce terme par des **contenus trompeurs et spécieux**, rendus **profondément crédibles** grâce à l'intelligence artificielle.

Mais ce terme peut également être directement inspiré du pseudonyme d'un utilisateur du site communautaire et social **Reddit** "**u/deepfake**" qui fut le premier à publier en novembre **2017** des **vidéos pornographiques** dans lesquelles il réussit à remplacer le visage des **actrices X** par celui de célébrités américaines.

Comme il est nécessaire de disposer d'une assez grande quantité d'images, au départ, pour que l'apprentissage soit vraiment performant, la médiatisation de certaines personnalités hollywoodiennes a malheureusement constitué une base de données privilégiée pour ce faussaire. Ainsi, si la jeune actrice **Daisy Ridley** fut la première victime de ses hypertrucages, le phénomène prit une réelle ampleur peu de temps après avec la **fausse sextape** mettant en scène **Gal Gadot**, (connue pour son interprétation de Wonder Woman).

La presse américaine (New York Times, Washington Post, Guardian, etc.) s'est rapidement inquiétée de la popularité massive de ces courtes séquences pornographiques, qui ne se cantonnent plus aujourd'hui à la fabrication de contenus obscènes mais deviennent au contraire de véritables **armes de communication politique**.

Que dit le cadre légal...

Ces trucages numériques étant considérés comme des « **infox** » peuvent relever de **la loi du 22 décembre 2018** relative à la **lutte contre la manipulation de l'information**, qui encadre la production et la diffusion de fausses informations.

Cependant, essentiellement destinée à réguler les **campagnes électorales**, elle s'applique difficilement à la quasi-totalité des *deepfakes* qui sont de **nature sexuelle ou pornographique**.

Aussi pour cette catégorie d'hypertrucages peuvent être mobilisés :

- l'article 226-8 du Code pénal qui prévoit un an d'emprisonnement et 15 000 euros d'amende "le fait de publier, par quelque voie que ce soit, le montage réalisé avec les paroles ou l'image d'une personne sans son consentement, s'il n'apparaît pas à l'évidence qu'il s'agit d'un montage ou s'il n'en est pas expressément fait mention".

- l'article 226-4-1 du Code pénal qui prévoit que "le fait d'usurper l'identité d'un tiers ou de faire usage d'une ou plusieurs données de toute nature permettant de l'identifier en vue de troubler sa tranquillité ou celle d'autrui, ou de porter atteinte à son honneur ou à sa considération est puni d'un an d'emprisonnement et de 15 000 € d'amende".

Pour aller un peu plus loin...

Quelques références scientifiques :

BARRE Aurélie, La force intime des images, *Littérature*, Volume 199, n° 3, 2020, pp. 86-100.

BRONNER Gérald, *la démocratie des crédules*, PUF, 2013.

CAZALS François, CAZALS Chantal, *Intelligence artificielle. L'intelligence amplifiée par la technologie*, De Boeck Supérieur, 2020.

DELFINO Rebecca A., Pornographic Deepfakes : The Case for Federal Criminalization of Revenge Porn's Next Tragic Act, *Fordham Law Review*, n°88, Issue 3, 2019, URL : https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=5640&context=flr

DUCOL Claudine, Le scoop : entre le savoir et l'opinion, *Communication et langages*, n°111, 1997. pp. 4-18.

GUARNERA Luca, GIUDICE Oliver, BATTIATO Sebastiano, Fighting Deepfake by Exposing the Convolutional Traces on Images, *IEEE*, 2020, Volume 8, pp. 165085-165098.

LANGA Jack, Deepfakes, real consequences : Crafting legislation to combat threats posed by deepfakes, *Boston University Law Review*, Volume 101, n° 2, 2021, pp. 761-801.

SCHICK Nina, *Deepfakes, The coming infocalypse*, Grand Central Publishing, 2020.

SIEGEL Dennis, KRAETZER Christian, SEIDLITZ Stefan, DITTMANN Jana, Media Forensics Considerations on DeepFake Detection with Hand-Crafted Features, Journal of imaging, Volume 7, n° 108, 2021, p. 108.

TÉTU Jean-François, L'émotion dans les médias : dispositifs, formes et figures, *Mots. Les langages du politique*, n°75, 2004, pp.9-20.

WESTERLUND Mika, The Emergence of Deepfake Technology : A Review, *Technology Innovation Management Review*, Novembre 2019, URL : https://timreview.ca/article/1282