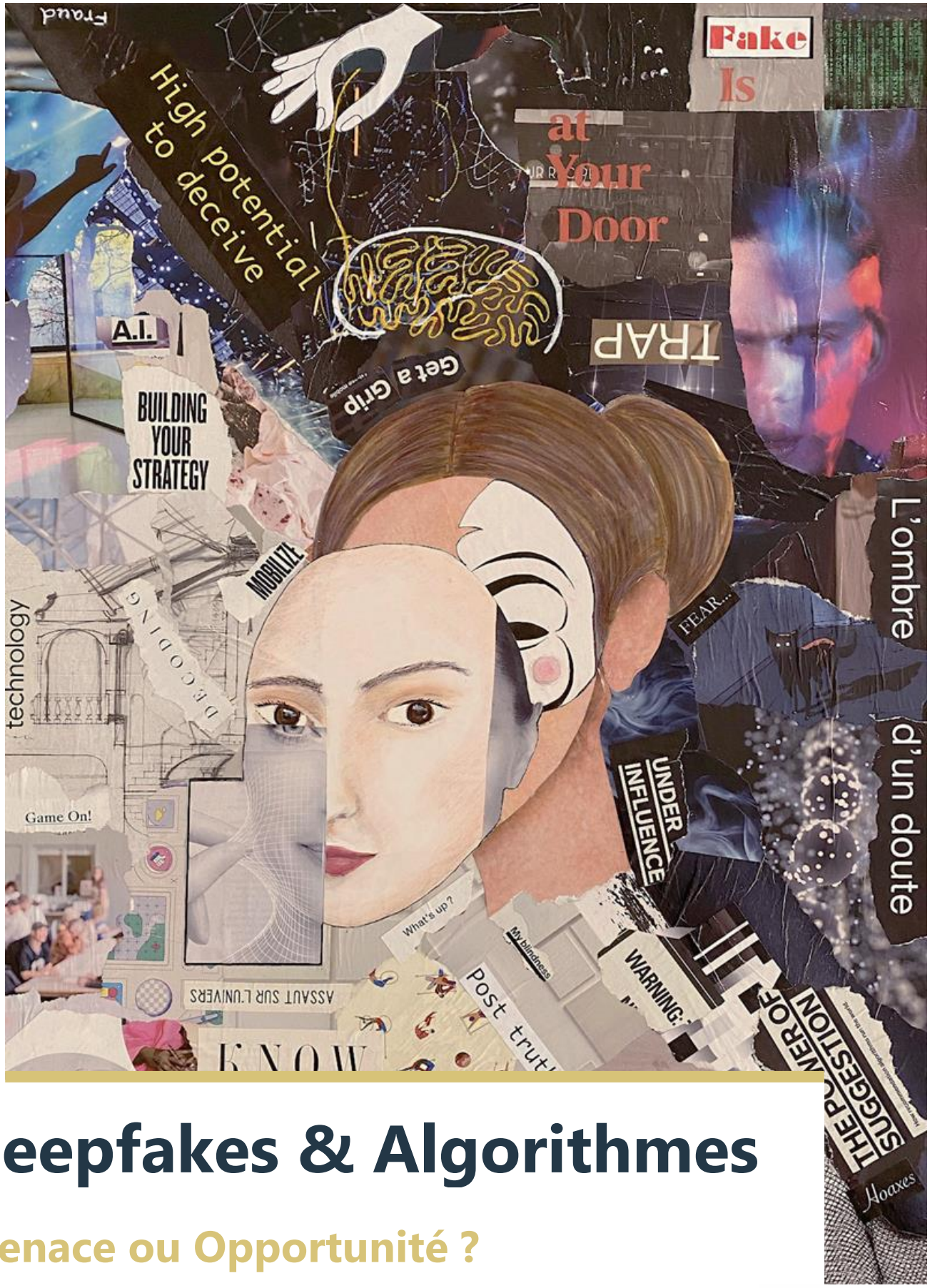


Deepfakes & Algorithmes

Menace ou Opportunité ?

Novembre 2020

(club) prAxis

Table des matières

Les auteurs **5**

Supervision 5

Groupe de travail 5

A propos du club Praxis **8**

Des prises de position pragmatiques et non partisanes 9

Un fonctionnement indépendant de toute subvention 9

Remerciements 10

Nos douze propositions pour repositionner l'Europe comme expert sur les algorithmes de génération de contenu et leader dans la lutte contre les *deepfakes* **11**

A. Faire de l'Europe un leader dans la lutte contre les *deepfakes* 11

B. Renforcer la responsabilité des plateformes au niveau européen 11

C. Construire un environnement réglementaire adapté à une lutte efficace contre les deepfakes 12

D. Protéger les citoyens de l'impact des deepfakes 13

Objectif du rapport **14**

Introduction **15**

Introduction aux deepfakes et aux algorithmes de génération de contenu **19**

Génération d'images 22

Des applications amusantes et créatives 22

Des applications plus inquiétantes 26

Génération audio 31

Génération de vidéos 33

Génération de texte 38

Open AI : GPT-2 et GPT-3 38

Génération de partitions 42

Génération de données 43

Contexte 46

Historique 46

Les algorithmes de génération de contenu 48

Démocratisation de la production des contenus 49

Les réseaux sociaux : diffusion accélérée, biais de vérité en mettant en avant les faux contenus 50

Qu'est-ce que l'éthique ? 52

Premières actions pour une éthique dans le net 52
Commission Européenne 53
OCDE 54
Rapport CNIL 56
Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires 57
Où en sont les plateformes internet 59
Facebook 60
Twitter 61
Google & Youtube 63
Microsoft 65
Plateformes asiatiques 67
Au-delà des acteurs de l'internet 70

Cadre légal, régulation et gouvernance 72
Cartographie des acteurs 72
Cadre juridique 72
Quels outils le droit français offre-t-il aujourd'hui pour faire face à des algorithmes de génération de contenu attentatoires aux droits des personnes ? 78
Cas n°1 : deepfake attentatoire au droit à l'image 79
Cas n°2 : deepfake diffamatoire (articles 29 et suivants de la loi du 29 juillet 1881 sur la liberté de la presse) 80
Cas n°3 : deepfake injurieux (articles 33 et suivants de la loi du 29 juillet 1881 sur la liberté de la presse) 80
Cas n°4 : deepfake diffuseur de fausses nouvelles (article 27 de la loi du 29 juillet 1881 sur la liberté de la presse et loi "anti fake news" du 22 décembre 2018) 81
Quelle doit être la responsabilité encourue par les plateformes ? 84
Les projets en cours 90
Exemples internationaux 92
En Europe 93
En dehors de l'Europe 93
Nos propositions 94

Comment reconnaître un contenu généré par un algorithme ? Comment aller plus loin pour se protéger du fake content ? 98
Détection visuelle de fausses images, quelques exemples ludiques 99
Détection algorithmique 105
Deepfake Detection Challenge 105
Détection algorithmique de faux textes 106

Recherche universitaire 108
Acteurs du domaine privé 112
Outils pour protéger le vrai 113
Fact-checking 99
Watermarks 113
Signature numérique 116
Blockchain 117
Sensibilisation et éducation 121
Rôle des auteurs 123
Rôle des plateformes 124

Conclusion 128

Annexes 129
Annexe n°1 : Deepfake Detection Challenge (DFDC) 129
Annexe n°2 : Les sept principes éthiques de la Commission Européenne en matière d'intelligence artificielle 132
Annexe n°3 : Commission Européenne, Livre blanc sur l'intelligence artificielle. Une approche axée sur l'excellence et la confiance, 2020 136
Annexe n°4 : Approche et conclusion du rapport de la CNIL, 2017 140
Annexe n°5 : Introduction à l'IA 143
Annexe n°6 : Fonctionnement des algorithmes de création de contenu 153

Les auteurs

Supervision

Emilie Rannou est diplômée de l'ENSAE et titulaire d'un master 2 en mathématiques appliquées de l'Université Paris Dauphine. Senior Manager chez Ekimetrics, elle conseille différentes entreprises sur leur stratégie data science et encadre des équipes de consultants data scientists pour implémenter et optimiser de nouveaux modèles. Emilie a travaillé 12 ans dans l'industrie tech et financière, notamment en tant que Executive Director chez JPMorgan Chase à New York et chez Criteo en tant que Senior Product Manager.

Groupe de travail

Alexis Benichoux est docteur en traitement du signal de l'Université de Rennes. Il a travaillé autour des méthodes d'apprentissage appliquées au traitement du signal sonore. Alexis est également diplômé du master d'apprentissage de l'ENS Cachan. Actuellement ingénieur de recherche chez Yubo, il développe des algorithmes de reconnaissance d'image et de la recommandation de contenu.

Rémi Forgeas est responsable du France Business Services Center pour un cabinet d'audit et de conseil international. Basé aux Etats-Unis depuis plus de vingt ans, il assiste des groupes français dans leur développement aux Etats-Unis. Il est diplômé de Sciences Po Paris et possède une maîtrise en économétrie.

Simon Gaillard est consultant en protection des données. Basé à New York, il est spécialisé dans le respect de la confidentialité en ligne et aide les entreprises sur leur stratégie de mise en conformité avec le RGPD et le CCPA. Diplômé de Sciences Po Paris, il conseille aussi les organisations internationales sur les politiques publiques, les nouvelles technologies et les questions de protection des données.

Jérémie Mary est docteur en informatique de l'Université Paris-Sud. Il est maître de conférences, habilité à diriger des recherches à l'université de Lille et est actuellement en détachement en tant que chercheur chez Criteo. Il est membre associé de l'équipe Inria SequeL qui est spécialisée dans la prise de décision séquentielle. Il a gagné et organisé 3 challenges pour les systèmes de recommandation dans des conférences majeures de machine learning utilisant des données Yahoo!, Adobe et Twitter (ICML'11, ICML'12, RecSys'14). Son travail de recherche actuel est orienté autour des séquences de recommandations et l'utilisation des modèles génératifs pour les systèmes de recommandation.

Minh Trinh est diplômé de l'Ecole Polytechnique et docteur en économie de l'Université Harvard. Il est consultant basé à New York et est spécialisé dans les méthodes quantitatives et de machine learning appliquées à l'économie et la finance. Il est l'auteur d'un ouvrage sur la gestion des projets en intelligence artificielle. Il est membre de l'American Finance Association et du CFA Institute.

Gabriel Turinici est un ancien élève de l'ENS Ulm et membre de l'Institut Universitaire de France. Il a été chargé de recherche en mathématiques et informatique à l'INRIA jusqu'en 2005 et exerce comme professeur à l'Université Paris Dauphine - PSL. Il est spécialisé dans les algorithmes de simulation, avec des applications en physique quantique, économie, finance et médecine. Il est l'auteur de plus d'une centaine de publications scientifiques et membre du comité de rédaction de plusieurs journaux.

Emilie Waxin est avocate au Barreau de Paris depuis 2010. Après avoir exercé pendant 8 ans en France et au Luxembourg, elle a créé sa propre structure, WE Avocats, en juin 2017. Elle exerce tant en conseil qu'en contentieux en droit des affaires et plus particulièrement en droit social, droit commercial et droit de la propriété intellectuelle. Elle enseigne également le contentieux au sein de l'Ecole de Formation du Barreau de Paris (EFB) et fait partie de plusieurs associations comme l'AFFJ (Association Française des Femmes Juristes).

A propos du club Praxis

Le Club Praxis est un think-tank franco-américain créé en 2007, sous l'impulsion de plusieurs dirigeants français, dont Henri de Castries. Il assume sa position d'outsider dans la vie publique française. Il cherche à promouvoir des idées neuves et innovantes pour le renouveau des institutions et du système économique sur la base des principes : stabilité, équité, transparence, recherche du meilleur échelon de décision politique au plus près du citoyen, éducation permanente et indépendance financière.

Des prises de position pragmatiques et non partisanes

Indépendant de toute école de pensée, le Club Praxis se veut le champion d'un dialogue renouvelé et sans a priori entre les citoyens et les décideurs politiques, notamment par le biais d'une plus grande transparence des données publiques et des processus de décision. À travers des propositions de réformes concrètes et actionnables, il souhaite remettre l'intérêt général à long terme au cœur de l'action publique, et contribuer à définir une nouvelle trajectoire pour une France à la pointe sur les sujets de recherche et plus ouverte sur le monde.

Un fonctionnement indépendant de toute subvention

Le Club Praxis fonctionne sur une base complètement bénévole, ne reçoit pas de subventions publiques ou privées. Il compte essentiellement des chercheurs, des économistes, des hauts fonctionnaires, des cadres supérieurs et des juristes qui souhaitent mettre leurs compétences et leur expérience au service de leur pays. Ce rapport n'engage que les auteurs.

Remerciements

Les auteurs remercient les personnes ayant fait part de leurs suggestions, notamment Aurélie Boris, Yann Coatanlem, Xavier Dupré, Eric Nahon, Adélaïde Sorel, Guy Sorman, Pablo Winant, le cabinet d'avocats Bold et la société de conseil en data science Ekimetrics pour leur contribution au rapport. Merci également à Vannina Malekzadeh pour son œuvre qui illustre le rapport.

Nos douze propositions pour repositionner l'Europe comme expert sur les algorithmes de génération de contenu et leader dans la lutte contre les *deepfakes*

Nous choisissons 12 recommandations pour faire de l'Europe un expert des algorithmes de génération de contenu et un leader dans lutte contre les *deepfakes*, définissant des standards de gouvernance et régulation autour des *deepfakes*, et protégeant ses citoyens, entre autres en les éduquant aux nouveaux défis créés par les *deepfakes* liés notamment à la manipulation, à la criminalité et au droit à l'image.

A. Faire de l'Europe un leader dans la lutte contre les *deepfakes*

1. **Soutenir l'amélioration et le développement des techniques de classification de contenu**, vrai, faux ou suspect (*fact-checking*, *watermarks*, algorithmes, *blockchain*), tout en définissant des standards européens, pour accompagner la population dans l'identification de contenu fiable.
2. **Investir dans la recherche scientifique sur les algorithmes de génération de contenu**, pour être mieux à même de les combattre mais aussi être capable de les utiliser pour des applications positives pour la société.
3. **Favoriser les partenariats européens**, mutualiser les ressources pour se donner les moyens de rattraper l'avancée de la Chine, de la Russie et des Etats-Unis dans ce domaine.

B. Renforcer la responsabilité des plateformes au niveau européen

4. **Renforcer la responsabilité encourue par les plateformes** (par exemple au sein du Digital Services Act). Imposer à chaque plateforme de mettre en place une structure de gouvernance interne sur la régulation de son contenu. Mettre chaque organe de gouvernance sous supervision européenne en s'inspirant de la régulation mise en place dans le domaine bancaire à la suite de la crise de 2008.

5. **Inciter les plateformes à mettre en place des outils pour diminuer l'impact des *deepfakes***

   Des sanctions telles que :

   - Une pénalité pour les auteurs publiant du contenu faux ou malveillant de manière non explicite.
   - Une mention claire mettant en garde l'utilisateur que la publication n'a pu être vérifiée, voire même qu'elle est suspecte.
   - Une limitation du nombre de destinataires d'une publication.
   - Une limitation des transferts d'une publication par utilisateur.

6. **Imposer aux plateformes de partager avec les autorités judiciaires** les informations qu'elles détiennent sur les auteurs présumés de *deepfakes*, dans le cadre d'enquêtes pénales, sans que les plateformes ne puissent leur opposer de motif légitime de refus.

C. Construire un environnement réglementaire adapté à une lutte efficace contre les *deepfakes*

7. **Favoriser l'accès au droit des victimes de *deepfakes***, d'usurpation d'identité ou de harcèlement (aux voies de recours et à l'information de leurs droits) par le développement et la visibilité des structures d'accompagnement (d'accueil physique ou plateformes étatiques).
8. **Renforcer les sanctions pénales existantes** contre les auteurs de contenus de *deepfakes* et les réévaluer en prenant en compte le nouvel environnement technologique et digital où la création de *deepfakes* est rendue possible.
9. **Renforcer au niveau étatique mais surtout supra-étatique (la transmission d'informations ne connaissant pas de frontières) la responsabilité des plateformes.** Dans cette voie :

- **Développer une organisation supranationale de cyber-coopération policière et judiciaire centrée sur la matière afin de faciliter les échanges d'informations entre les Etats.**
- Permettre dans certains cas exceptionnels, **lorsqu'ils diffusent un contenu clairement illégal, d'ordonner la fermeture définitive d'un site**, d'un profil, d'un compte, d'une page, d'un groupe, sans que le site attaqué puisse se défendre
- **Limiter et quantifier légalement le délai de retrait d'un contenu illicite** signalé aux plateformes sous peine de sanctions financières lourdes
- Obliger les plateformes à **vérifier** les données d'identification lors de la création d'un compte et, en **l'absence d'identification réelle de l'auteur, fermer le compte sans délai**
- Imposer aux hébergeurs situés hors de France de désigner un représentant en France qui assume la responsabilité de l'hébergeur qui n'aurait pas retiré dans ce délai légalement fixé un contenu signalé comme illégal
- Assortir systématiquement les sanctions financières infligées aux plateformes d'une mesure de publicité

D. Protéger les citoyens de l'impact des *deepfakes*

10. **Faire de la lutte contre les *deepfakes* un objectif européen pour 2021.** Mettre en place des campagnes de sensibilisation pour expliquer de manière pédagogique les *deepfakes*, leurs utilisations et les risques associés en termes de détournement des images, de la voix et des vidéos, tout en insistant sur les applications possibles en manipulation d'opinion.

11. **Renforcer la responsabilité encourue par les auteurs**. Imposer à chaque auteur de signaler tout contenu *deepfake*.

12. **Impliquer les employeurs dans la lutte contre les *deepfakes*.** Les inciter à adopter des chartes éthiques, à informer leurs employés, clients, fournisseurs et partenaires de leurs bonnes pratiques éthiques sur le sujet. Encourager la mise en place de formations autour des *deepfakes*.

Objectif du rapport

Dans le présent rapport, nous nous penchons spécifiquement sur les **algorithmes de génération de contenu**. De quoi sont-ils capables aujourd'hui ? Comment fonctionnent-ils ? Dans quelle mesure savons-nous reconnaître un contenu produit par un algorithme d'intelligence artificielle ? Que sont les ***GANs***, *Generative Adversarial Networks*, algorithmes qui se cachent derrière les *deepfakes* les plus sophistiqués ? Sont-ils uniquement voués à réaliser des *deepfakes* ou peut-on identifier des applications bénéfiques pour la société ? Comment sensibiliser la population pour mieux reconnaître un *deepfake* et déjouer les tentatives de manipulation ? Quel cadre légal est en place autour de ces sujets et comment devrait-il évoluer pour s'adapter aux nouveaux algorithmes ? Nous allons aborder de multiples questions éthiques liées à ce sujet mais aussi explorer comment il est possible de mieux évoluer dans un environnement où les algorithmes de génération de contenu sont de plus en plus présents.

Notre objectif est, dans un premier temps, pédagogique : **volonté de vulgarisation du sujet pour les pouvoirs publics et les entreprises**, et de faire prendre conscience des enjeux vertigineux que les algorithmes de génération de contenus soulèvent. Dans un deuxième temps, il s'agit de **peser sur le débat autour du cadre numérique, réglementaire et légal** des *deepfakes* ainsi que sur le débat économique et social, français et européen, pour mieux sensibiliser la population à ces enjeux.

C'est dans cette optique que nous, penseurs, chercheurs et experts dans les domaines scientifiques, légaux, éthiques, tous passionnés par ce sujet des algorithmes de génération de contenus et par les *deepfakes*, nous sommes réunis et que nous vous présentons ce rapport, à un moment préoccupant de l'histoire, où manipuler du contenu paraît de plus en plus facile et où le pouvoir d'influence des *leaders* numériques sur les opinions et les élections ne cesse de croître. Il nous apparaît urgent de prendre conscience, à plus grande échelle, que les algorithmes de génération de contenu ont ouvert une véritable boîte de Pandore. Développer une connaissance plus approfondie dans ce secteur sera essentiel pour positionner l'Europe comme le leader dans le combat contre les *deepfakes*.

Introduction

En mai 2019, Facebook est sous le feu des critiques : le réseau social a laissé circuler une **fausse vidéo** de la présidente de la Chambre des représentants des États-Unis, Nancy Pelosi, dans laquelle elle semblait ivre. Un trucage rendu possible par un ralentissement de la voix tout en gardant un visuel crédible. Devenue virale, la vidéo a été vue par plus de 2 millions de personnes en deux jours entraînant une vague de réactions sur les réseaux sociaux. Cette vidéo a également été partagée par des proches du président Républicain Donald Trump, en période de fortes tensions entre Nancy Pelosi et le président américain, chacun cherchant à décrédibiliser son adversaire. Face à la pression pour retirer la vidéo, Facebook invoque la **liberté d'expression.** Ne s'estimant ni garant ni juge de la véracité des contenus mis en ligne, l'entreprise annonce son refus de supprimer cette vidéo bien qu'elle soit trompeuse et susceptible de manipuler l'opinion. Facebook propose néanmoins **d'étiqueter la vidéo de la mention : « partiellement fausse ».**

Le débat entre manipulation et liberté d'expression n'est pas nouveau. Il prend cependant une **ampleur considérable** avec, d'une part, le pouvoir croissant des réseaux sociaux qui permettent de toucher un nombre accru de personnes, et d'autre part, avec les progrès technologiques qui **favorisent la création de faux contenus** et notamment des ***deepfakes*** : faux contenus créés à l'aide de technologies nommées *deep learning* ou apprentissage profond. Toutefois, Facebook change de position début 2020 à la suite de la mise en scène de son fondateur et Directeur Général Mark Zuckerberg, dans une vidéo truquée où son personnage annonce qu'il est prêt à se servir des milliards de données volées aux citoyens pour contrôler le futur et qu'il doit tout à une organisation secrète nommée "Spectre". Bien que la vidéo, œuvre de deux artistes *hacktivistes,* soit de piètre qualité et le subterfuge manifeste, cela a-t-il pu jouer dans la balance pour convaincre Mark Zuckerberg du pouvoir nocif des *deepfakes* ? Quoi qu'il en soit, en janvier 2020, **Facebook publiera un communiqué les interdisant**.

Les *deepfakes* apparaissent désormais comme un outil de **manipulation** voire de **propagande** dont **l'impact sur la société est encore peu maîtrisé.** Leur existence et utilisation soulèvent de nombreuses questions d'ordre **légal** et **éthique** mais, au-delà des lois et règles de gouvernance

susceptibles d'être mises en place, un problème fondamental demeure, celui de **l'incapacité à détecter un *deepfake***. Alors que la technologie évolue, **il s'avère de plus en plus compliqué d'identifier un faux**. Développer une **connaissance européenne** de ces technologies, déjà maîtrisées par de nombreux experts notamment en Chine et en Russie, apparaît donc urgent pour mieux les appréhender et les encadrer.

Le rapport Villani, publié en 2018, détaille pourquoi **la France doit impérativement investir plus dans le domaine de l'Intelligence Artificielle** (IA). Depuis des dizaines d'années, l'intelligence artificielle nourrit craintes et fantasmes, inspirant nombre de réalisateurs et écrivains mais son impact sur la société était jusqu'à récemment encore ténu. En effet, l'intelligence artificielle ne paraissait pas encore en mesure d'inquiéter l'humanité, ne reproduisant alors que ce que l'homme lui avait appris. Un premier événement marque pourtant les esprits en 1997 : l'ordinateur *Deep Blue* (IBM) bat le champion du monde d'échecs Gary Kasparov. L'événement bouscule une société hésitante : doit-elle se préoccuper de la révolution imminente et néanmoins prévisible qui se profile ? Finalement, c'est surtout la forte progression de la puissance de calcul des ordinateurs des 10 dernières années qui remet les algorithmes d'intelligence artificielle sur le devant de la scène. En effet, le temps d'apprentissage était jusque-là l'obstacle majeur auquel se heurtait systématiquement l'IA. Grâce aux **améliorations des puissances de calcul**, il est *de facto* réduit. Le second problème était lié aux données disponibles à cet apprentissage. Cette fois, c'est l'avènement du numérique qui en a eu raison, du moins en partie, grâce à la prolifération des applications et à la capacité à collecter, stocker et partager massivement des données, notamment à l'aide du *cloud*.

Il nous apparaît donc primordial de développer une meilleure connaissance de ces technologies auprès du public, notre objectif principal étant d'**exposer leur fonctionnement** de manière pédagogique et de **proposer des suggestions** sur leur utilisation et leur **encadrement**. C'est pourquoi, nous commencerons par montrer dans quelle mesure leurs domaines d'application sont riches et variés, notamment dans la **génération de textes, d'images, de sons et de vidéos**, en donnant des exemples concrets d'applications nocives mais aussi bénéfiques pour la société. Nous nous pencherons, ensuite sur les **enjeux éthiques**. Nous étudierons dans une troisième temps la

réglementation encadrant les *deepfakes* et comment nous envisageons son évolution. Finalement, notre volonté étant de protéger la population, nous avons présenté les **pistes existantes pour limiter l'impact négatif des *deepfakes***.

1. Introduction aux *deepfakes* et aux algorithmes de génération de contenu

Il faut, en général, plusieurs années pour qu'un domaine de Recherche et Développement trouve des applications concrètes et utiles pour les citoyens. Ce n'est pas le cas pour les algorithmes de génération de contenu car trois ans après leur création en 2017, les *GANs* sont déjà largement utilisés par le grand public et leur principale application, les *deepfakes* commencent à faire parler d'eux. Pour mesurer l'intérêt que suscite un sujet, une approche simple est de regarder les statistiques de Google. Vous trouverez ci-dessous un extrait des statistiques des recherches sur le deepfake dans le monde.

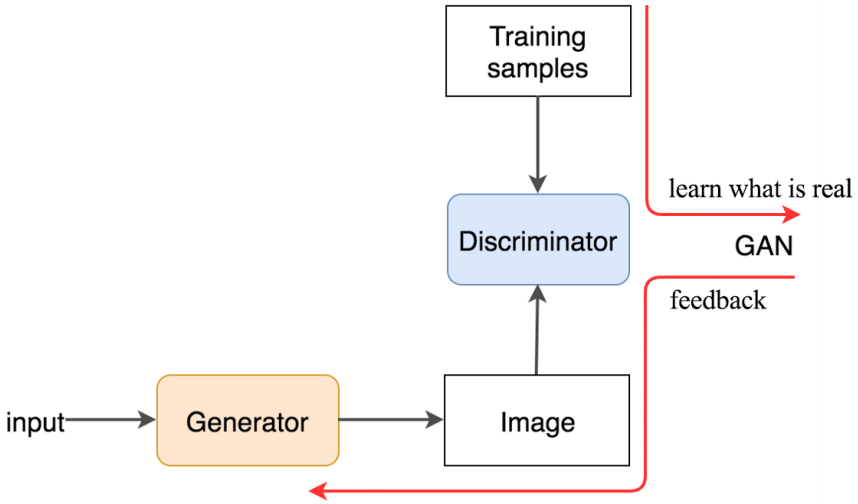
Source : Google Trends

Le premier pic de recherches détecté par Google Trends date de Janvier 2018. Il correspond aux scandales de *deepfake porn* (images et vidéos pornographiques créées par des algorithmes de *deep learning*). L'intérêt pour les *deepfakes* baisse ensuite rapidement mais remonte progressivement en 2019. C'est en effet en 2019 que sont publiés deux *deepfakes* qui ont créé le buzz : l'un met en scène Barack Obama dans lequel il insulte Donald Trump. L'autre met en scène Mark Zuckerberg dans lequel il avoue avoir compris l'intérêt de voler les données des citoyens grâce à une organisation secrète nommée *Spectre*. C'est aussi en septembre 2019 que Facebook lance le *Deepfake Detection Challenge* dont nous reparlerons.

Mais peu sont ceux qui appréhendent bien ce qu'est un *deepfake*.

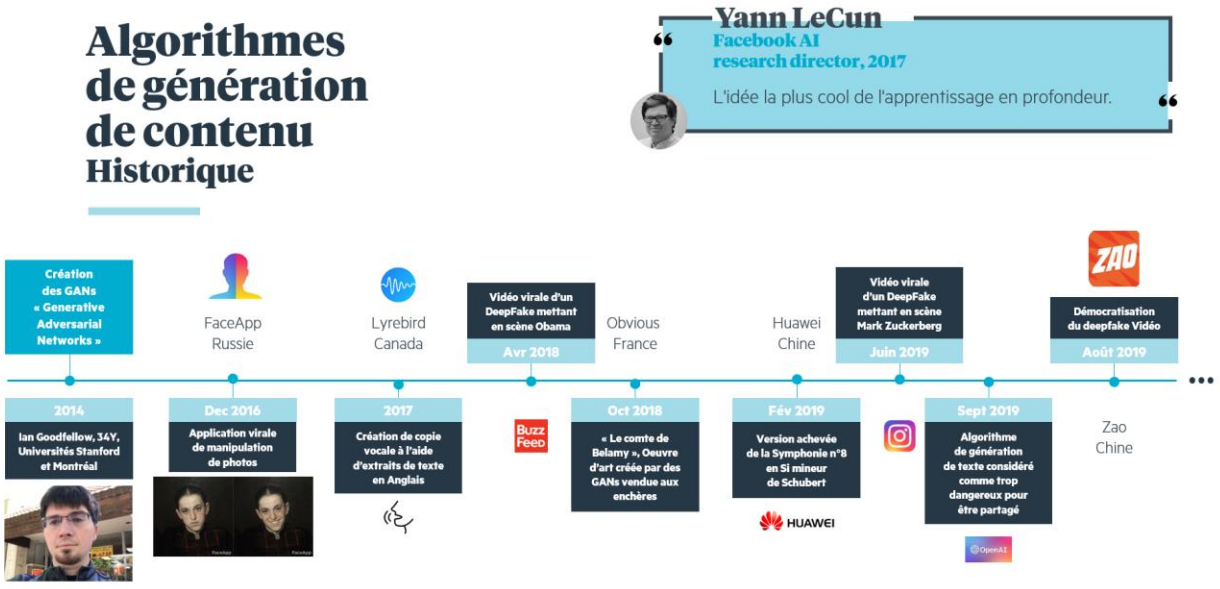
Qu'est-ce qu'un *deepfake* ? C'est un faux contenu généré par des algorithmes d'apprentissage profond, le *deep learning* en anglais, d'où le nom de *deepfake : deep* pour profond et *fake* pour faux. Mais qu'est-ce que le *deep learning* ? Tout d'abord, il faut introduire ce qu'on appelle un réseau de neurones. C'est un algorithme d'intelligence artificielle qui s'inspire du fonctionnement du cerveau et notamment du transfert d'information entre les neurones via les synapses. Une couche simple de neurones permet à l'algorithme d'apprendre des règles simples et de classifier par exemple un mot en féminin, masculin ou en neutre à l'aide d'un apprentissage sur un dictionnaire. Un réseau de neurones avec plusieurs couches permettra un apprentissage plus complexe, comme l'apprentissage d'un concept, et permettra d'identifier si une image est celle d'un chat, d'un chien ou d'un singe à l'aide de l'apprentissage sur des images similaires. Ces réseaux avec plusieurs couches sont ce que nous nommons les réseaux profonds, et le *deep learning* regroupe l'ensemble des algorithmes élaborés à l'aide de réseaux profonds de neurones.

Nous allons nous pencher dans ce rapport plus en détail sur les algorithmes phares utilisés pour la création de *deepfakes* : les *GANs* (*Generative Adversarial Networks*). Le fonctionnement de ces algorithmes de génération de contenu, dont nous donnerons plus de détails dans la partie technique, est de confronter deux réseaux profonds (*deep networks*) : un réseau générateur (*generator*) et un réseau discriminateur (*discriminator*). Pour prendre un exemple concret, plaçons-nous dans un cas où nous souhaitons générer une photo d'une voiture avec le style du peintre Gauguin. Le réseau générateur transforme la photo initiale de la voiture et génère une image légèrement modifiée à l'aide de ce qu'il a appris des tableaux de Gauguin. Cette image générée, laquelle n'est donc pas une image réelle, va être, dans un deuxième temps, confrontée par le discriminateur qui, comme un inspecteur, pourra précisément évaluer l'image produite par le générateur avec l'objectif de conclure qu'elle est fausse ou réelle. S'il conclut qu'elle est réelle, le discriminateur apprend alors qu'il s'est trompé et améliore sa connaissance. S'il conclut qu'elle est fausse, le générateur apprend que sa création n'était pas assez bonne et améliore également sa connaissance, ce qui lui permet de proposer une nouvelle copie au discriminateur. Et ainsi de suite... Ce processus permet au générateur de produire des données de plus en plus réalistes.



Les algorithmes de génération de contenu ont été introduits en 2014 par Ian Goodfellow, alors chercheur à l'Université de Montréal (il a également obtenu un master en informatique à l'université Stanford). Il travaille actuellement chez Apple, après un poste chez Google Brain et une collaboration dans le cadre d'Open AI, entreprise à but non lucratif plafonné, fondée par un collectif dont Elon Musk fait partie.

Voici un historique des algorithmes de génération de contenu :



Nous allons détailler ces nombreuses applications par la suite.

1.1. Génération d'images

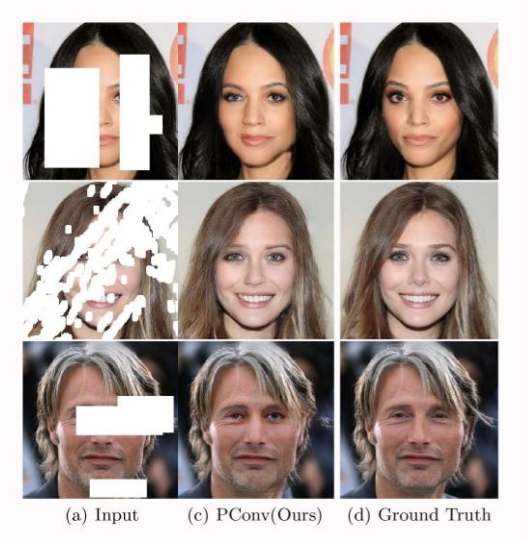
1.1.1. Des applications amusantes et créatives

Tout possesseur de smartphone peut aujourd'hui s'amuser à mettre des filtres pour styliser ses photos. En effet, entre autres et ce n'est pas la seule, l'application Deep Art Effects permet de générer des images dans le style de grands peintres tels Monet ou Van Gogh. Peu sont ceux qui soupçonnent que derrière ces filtres, se cache une des dernières technologies de machine learning et notamment une catégorie particulière de *GANs* (CycleGAN[1]). Par exemple, nous voyons ci-dessous une photographie stylisée grâce à l'apprentissage du style de nombreuses peintures existantes.

Source : Jun-Yan Zhu et al., 2017

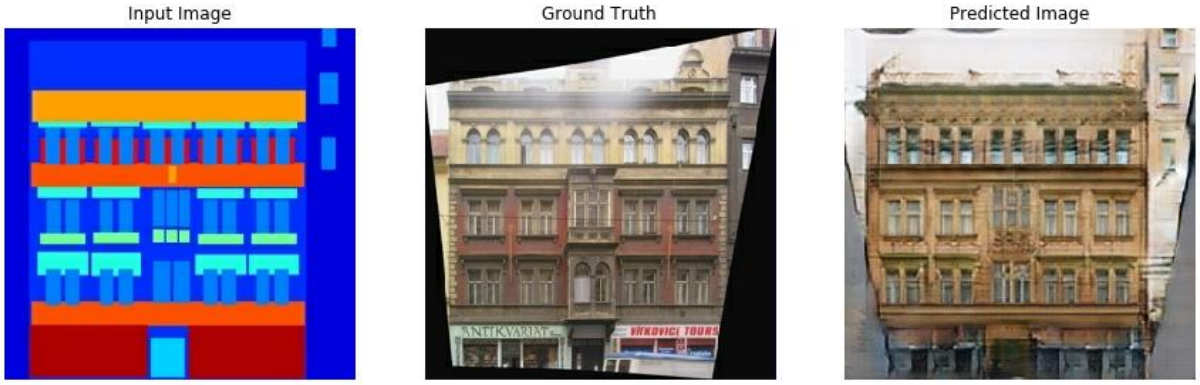
Les applications dans le domaine de l'image peuvent aller au-delà du simple divertissement et se révéler extrêmement utiles pour ceux qui cherchent à compléter une photo ou une fresque abîmée. Ci-dessous se présentent les résultats d'un algorithme implémenté en avril 2018 par l'entreprise américaine NVIDIA pour compléter des images dont certaines parties sont manquantes. La colonne (a) est la photo incomplète, la (d) est la réalité et la colonne (c) est la prédiction de l'algorithme. Bien que la prédiction soit très réaliste, il faut néanmoins nuancer ces résultats impressionnants et noter que l'algorithme a pu apprendre sur de nombreux exemples similaires.

[1] Jun-Yan Zhu, Taesung Park, Phillip Isola, et Alexei A. Efros, Un CycleGAN est un GAN qui utilise 2 générateurs et 2 discriminateurs. *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*, 2017, https://junyanz.github.io/CycleGAN



Source : NVIDIA, 2018

Les algorithmes permettent de générer des images à partir d'autres images mais sont aussi capables de créer des photos à partir de dessins et même juste de textes ! Ci-dessous un exemple de génération de photographies à l'aide d'un dessin (avec la librairie Pix2Pix[2]) :



Source : TensorFlow Pix2Pix

[2] Pix2Pix est un module de *Tensorflow*, librairie open source (développée par Google) très largement utilisée par la communauté des développeurs. *Tensor Flow*, *Pix2Pix*, https://www.tensorflow.org/tutorials/generative/pix2pix

Et ci-dessous un exemple de génération d'images à partir de textes (avec *StackGans*[3]) :

Source : Rajat Garg, 2019

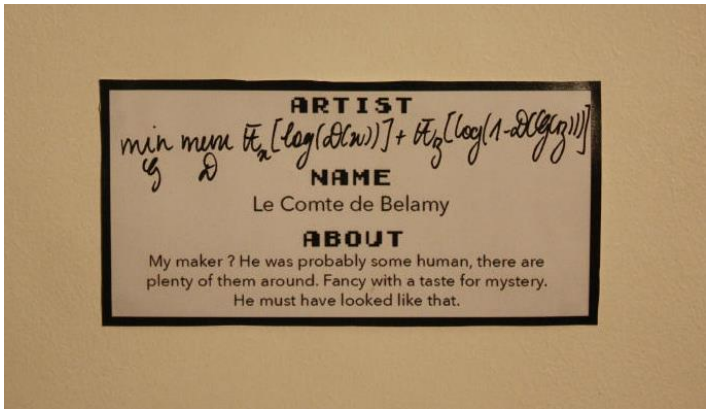
Passons maintenant à un exemple qui a marqué les esprits en 2018 : la première vente aux enchères d'un tableau généré par un algorithme ! Le pouvoir créatif de ces algorithmes et leur potentiel pour faire rêver et créer différemment a bien été compris par le collectif français *Obvious*[4] qui a vendu aux enchères, chez Christie's à New York en octobre 2018, le premier tableau généré par un algorithme.

Ce tableau, dont le titre est "Le Comte de Bellamy" en hommage à Ian Goodfellow, créateur des algorithmes de génération de contenu, a été conçu à l'aide de ces algorithmes, ayant effectué leur apprentissage sur environ 15 000 portraits peints entre le XVème et XIXème siècle. Le collectif a apposé comme signature du tableau une formule mathématique correspondant à l'apprentissage des algorithmes de génération de contenu. Il a été vendu chez Christie's pour plus de 400,000 USD[5].

[3] Rajat Garg en a développé un exemple. Rajat Garg, *Implementing StackGAN using Keras — Text to Photo-Realistic Image Synthesis,* 2019, https://medium.com/@mrgarg.rajat/implementing-stackgan-using-keras-a0a1b381125e
[4] Obvious, https://obvious-art.com
[5] CHRISTIE'S, *Is artificial intelligence set to become art's next medium?,* 2018, https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx

Source : CHRISTIE'S, 2018 © Obvious

Nous montrons ci-dessous un zoom de la signature, amusante, et qui n'appose non pas le nom du collectif mais la formule mathématique caractéristique des algorithmes de génération de contenu. Nous reviendrons à cet exemple dans la partie de ce rapport discutant de l'angle juridique et notamment de la propriété intellectuelle des contenus ou œuvres produites par ces algorithmes.



Source : CHRISTIE'S, 2018 © Obvious

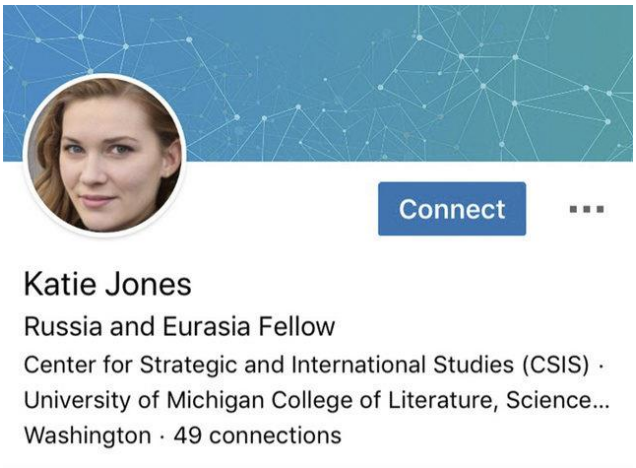
1.1.2. Des applications plus inquiétantes

Pour sensibiliser la population à l'impact des algorithmes de génération de contenu, un ingénieur chez Uber a créé un site : *thispersondoesnotexiste.com*[6] en février 2017 à l'aide d'un algorithme open source de NVIDIA. L'objectif était de montrer à quel point un visage généré par cet algorithme peut être réaliste. Ces visages peuvent malheureusement être utilisés pour créer de faux profils sur les réseaux sociaux. En effet, Linkedin, Facebook et Twitter font face à de grandes difficultés pour éviter la création de faux profils. Ces profils peuvent être créés de manière automatique et relayer massivement de fausses informations sur les réseaux. Notamment, ils peuvent aussi être créés pour influencer une note positive pour un restaurant (des entreprises pouvant "acheter" un certain nombre de likes ou de commentaires positifs), voire négative pour décrédibiliser un concurrent ou encore à des fins politiques pour mettre en avant ou décrédibiliser un candidat.

> ***Ces profils peuvent être créés de manière automatique et relayer massivement de fausses informations sur les réseaux.***

[6] This Person Does Not Exist, https://thispersondoesnotexist.com/

Pour montrer que la génération d'images peut être utilisée à des fins suspicieuses liées à la sécurité ou à la politique, l'agence de presse américaine, Associated Press, a conduit des enquêtes pour détecter des cas similaires. En effet, elle a pu repérer le faux profil Linkedin ci-dessous, dont la photo de profil a été générée par un algorithme. Selon une publication dans un journal : "*dans la quelque cinquantaine de connexions de Katie Jones, on retrouve un certain nombre de personnalités importantes à Washington, comme le haut conseiller d'un sénateur ou des sous-secrétaires d'État adjoints. La jeune femme aurait travaillé de nombreuses années au Centre d'études stratégiques et internationales à Washington, d'après son profil, et c'est sans doute pour cela qu'elle a pu ajouter autant de politiques dans ses contacts. Mais pourtant, celle-ci n'a jamais pu y travailler puisqu'elle n'existe tout simplement pas.*"[7]



Source : Rémi Lou, 2019

Des personnalités haut placées ont été bernées par un simple faux profil, ce qui peut soulever de nombreux problèmes de sécurité. On peut également citer dans l'actualité récente la désactivation par Facebook et Twitter de comptes suspectés par le FBI d'avoir été créés par l'Etat russe dans le

[7] Rémi Lou, *Des espions infiltreraient LinkedIn avec des photos de profils créées par l'IA*, 2019, https://www.journaldugeek.com/2019/06/14/des-espions-infiltreraient-linkedin-avec-des-photos-de-profils-creees-par-lia/

but de recruter des pigistes américains à des fins de déstabilisation. Les photos de profil de nombreux comptes auraient été créés synthétiquement par un algorithme[8]. Ce qu'a détaillé le chercheur Ben Nimmo dans l'image ci-dessous.

Source : Twitter

Sur un autre plan, dès lors que la génération de contenu est utilisée à partir de photos humaines, le sujet devient compliqué et les débats éthiques sont nombreux.

L'application FaceApp[9], développée en 2017 par la compagnie Russe Wireless Lab, était selon le Huffington Post, en tête des classements des applications les plus téléchargées sur iPhone et Android le 17 juillet 2019. En partant d'une photo du visage de l'individu, elle offre la possibilité de voir ce visage en état vieilli. Par exemple, le cuisinier Gordon Ramsey s'est prêté au jeu.

[8]Le Monde, *Facebook et Twitter bloquent une opération d'influence reposant sur un « média indépendant » attribuée à la Russie*, 2020 https://www.lemonde.fr/pixels/article/2020/09/02/facebook-et-twitter-bloquent-une-operation-d-influence-reposant-sur-un-media-independant-attribuee-a-la-russie_6050716_4408996.html

[9]FaceApp utilise des algorithmes IA mais plus simples que les *GANs*. Wikipédia, *FaceApp*, https://en.wikipedia.org/wiki/FaceApp,

Source : Gordon Ramsey's account

En plus du vieillissement, l'application FaceApp propose d'autres fonctionnalités, comme changer un homme en femme ou rajouter un sourire. En 2017, l'application a rajouté un filtre pour rendre une photographie sexy avec le filtre "hot." FaceApp s'est vue fortement critiquée, les notions de beauté étant totalement subjectives. Il est délicat de demander à un algorithme de définir ce qui est plus sexy : ici, le filtre "hot" blanchissait systématiquement les peaux des hommes noirs :

Source: FaceApp

Accusée de racisme et à la suite de plusieurs critiques comme " *FaceApp n'est pas simplement nulle, c'est raciste. Le filtre hot a blanchi ma peau et modifié mon nez pour le rendre plus européen* ", " *FaceApp nous apprend que le blanc c'est beau. #racisme* " ou " *Pourquoi beau veut-il dire blanc ?* ", FaceApp a supprimé ce filtre et le fondateur Yaroslav Goncharov a également présenté ses excuses.

FaceApp a aussi fait polémique à cause de l'utilisation des données personnelles : les utilisateurs devaient consentir à céder leurs droits sur leurs photos de manière "irrévocable" et "permanente" :

> You grant FaceApp a perpetual, irrevocable, nonexclusive, royalty-free, worldwide, fully-paid, transferable sub-licensable license to use, reproduce, modify, adapt, publish, translate, create derivative works from, distribute, publicly perform and display your User Content and any name, username or likeness provided in connection with your User Content in all media formats and channels now known or later developed, without compensation to you. When you post or otherwise share User Content on or through our Services, you understand that your User Content and any associated information (such as your [username], location or profile photo) will be visible to the public.
>
> You grant FaceApp consent to use the User Content, regardless of whether it includes an individual's name, likeness, voice or persona, sufficient to indicate the individual's identity. By using the Services, you agree that the User Content may be used for commercial purposes. You further acknowledge that FaceApp's use of the User Content for commercial purposes will not result in any injury to you or to any person you authorized to act on its behalf. You acknowledge that some of the Services are supported by advertising revenue and may display advertisements and promotions, and you hereby agree that FaceApp may place such advertising and promotions on the Services or on, about, or in conjunction with your User Content. The manner, mode and extent of such advertising and promotions are subject to change without specific notice to you. You acknowledge that we may not always identify paid services, sponsored content, or commercial communications as such.

Ceci a créé beaucoup d'incertitudes quant à l'utilisation des données récupérées d'autant plus que la société est Russe et son siège établi à Saint-Pétersbourg.

D'autres applications sont plus directement nocives, comme par exemple le cas de l'application *DeepNude*, qui n'existe plus officiellement aujourd'hui.

Source : DeepNude App

Cette application, officiellement fermée quelques semaines après son lancement en ligne le 23 juin 2019, mais qui est en fait toujours téléchargeable en ligne, permet de dénuder une femme grâce à l'IA à partir d'une simple photo. Créée pour le divertissement des utilisateurs, cette application porte préjudice aux femmes dont la photo a été utilisée, porte atteinte à leur pudeur et les humilie notamment lorsque les fausses photos dénudées sont publiées sur le web. L'application a officiellement fermé, en invoquant pour raison principale son succès qui a été bien plus important que prévu, et n'ayant pu faire face au trafic de demandes d'utilisation. DeepNude invoque aussi des questions éthiques liées à la forte probabilité d'utilisation à mauvais escient de l'application. Github, logiciel sur lequel les développeurs partagent le code de leurs algorithmes gratuitement, a pu observer de multiples codes répliquant le code de DeepNude et a tenté de les interdire pour cause de contenus obscènes. Cependant on trouve encore sur la plateforme de nombreuses versions de code de DeepNude[10].

1.2. Génération audio

De nombreuses années après avoir lancé son logiciel de retouche d'images *Photoshop*, la société américaine Adobe annonce travailler sur un logiciel de reconstitution de voix. Nommé VOCO (en référence à *Voice Conversion*), ce logiciel est un outil de synthèse vocale. Il permet de générer un discours à l'aide d'un texte et d'un échantillon de voix de vingt minutes, et ceci grâce à un algorithme de *text to speech*. Il est présenté en Novembre 2016[11], le présentateur présentant un

[10] GihHub, *deepnude*, https://github.com/topics/deepnude
[11] #VOCO https://www.youtube.com/watch?v=I3l4XLZ59iw&feature=youtu.be

discours d'origine "*I kissed my wife and my dogs*" et tapant différents textes dans le logiciel. Il permet de générer différentes phrases comme "I kissed my wife and my wife" et finalement "I kissed Jordan three times". Nous notons que la voix est effectivement la même mais la tonalité et le timbre de la voix ne semblent pouvoir être modulées. Ce projet est encore en version test et n'est pas encore commercialisé.

Un peu avant, en septembre 2016, Google DeepMind avait annoncé la création de Wavenet[12], son logiciel de création de son. Il est capable de reproduire n'importe quelle voix et n'importe quel son. Comme la voix d'une personne avec un accent Allemand par exemple, ou encore le son d'un piano. En 2018, Google annonce avoir pu remplacer une voix par une autre (*voice swapping*). En 2019, l'entreprise annonce qu'il suffit d'un échantillon de quelques minutes de voix pour générer un discours.

En 2017, la génération audio fait un bond en avant : la start-up Canadienne Lyrebird[13] a développé un algorithme permettant de synthétiser un discours à partir d'un échantillon de voix d'une minute. Elle est également capable de synthétiser des voix avec des émotions différentes. La start-up prévoit de nombreux domaines d'application comme les assistants personnels, la lecture de livres audio, les objets connectés, la synthèse de voix pour des personnes en situation de handicap, la réalisation de films ou de jeux vidéo, etc.

Sur son site internet, Lyrebird aborde les questions éthiques. En effet, un enregistrement audio peut servir de pièce à conviction dans un procès et influencer une décision. Lyrebird estime que son outil peut sensibiliser les auditeurs à ne pas considérer comme une preuve irréfutable un enregistrement audio[14] :

"*En révélant l'existence de cette technologie et en la rendant disponible, nous souhaitons conjurer l'apparition de tels risques. Nous espérons que le public prendra conscience qu'imiter une voix est*

[12] WAVENET: A GENERATIVE MODEL FOR RAW AUDIO, Sep.2016, https://arxiv.org/pdf/1609.03499.pdf
[13] Descript, *Overdub*, https://www.descript.com/lyrebird-ai?source=lyrebird
[14] Nelly Lesage, *Flippant : cet algorithme recrée la voix de quelqu'un à partir d'une minute d'enregistrement*, 2017, https://www.numerama.com/tech/251861-cet-algorithme-peut-recreer-la-voix-de-quelquun-a-partir-dun-enregistrement-audio.html

devenu possible et que cela doit relativiser à l'avenir la valeur de preuve accordée aux enregistrements audio."

La peur de telles applications se concrétise lorsqu'en septembre 2019, le Directeur Général d'une firme énergétique britannique se fait voler 220 000 dollars via un *deepfake* audio. Un individu a appelé le directeur général en utilisant la voix de son président pour lui demander un virement, celui-ci ne s'est pas méfié et a réalisé le virement demandé.

La peur de telles applications se concrétise lorsqu'en septembre 2019, le Directeur Général d'une firme énergétique britannique se fait voler 220 000 dollars via un deepfake audio.

Les technologies utilisées sont accessibles à tous, ce qui nous montre l'urgence de mettre en place des systèmes de sécurité efficaces pour faire face aux imitations de voix par des algorithmes, comme en mettant en place des *watermarks* (tatouages, filigranes). Ce dernier point sera détaillé par la suite.

1.3. Génération de vidéos

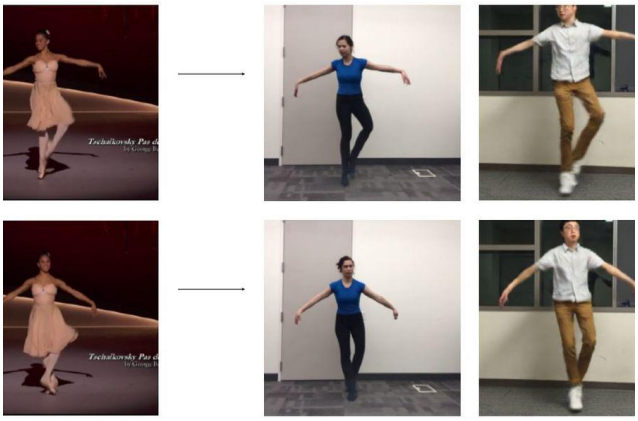
Exemples ludiques

Des chercheurs de l'université de Berkeley ont publié en août 2018 un article de recherche nommé *Everybody Dance Now*[15] associé à une vidéo postée sur Youtube[16]. Leur article de recherche détaille comment ils peuvent produire une vidéo représentant une personne 'cible' dansant de manière similaire à un danseur 'source' dont la vidéo a été donnée en entrée de leur algorithme. Ci-dessous, des images du rendu avec en source une danseuse étoile professionnelle et en cible deux individus ne sachant pas danser. L'algorithme est donc capable de transférer les mouvements

[15] Caroline Chan, Shiry Ginosar, Tinghui Zhou et Alexei A. Efros, *Everybody Dance Now*, 2019, https://arxiv.org/pdf/1808.07371.pdf

[16] La vidéo peut être consultée sur le lien suivant : https://youtu.be/PCBTZh41Ris

de la danseuse étoile sur le corps d'une personne cible préalablement filmée en train de réaliser des mouvements basiques.



Source : Caroline Chan et al., 2019

En mai 2019, Samsung lance une nouvelle application[17] qui permet d'animer un visage à partir d'une seule photo. L'algorithme, appuyé sur les *GANs*, permet l'animation d'un tableau comme celui de La Joconde ci-dessous :

Source: Bastien L., 2019

[17] Bastien L, *L'IA de Samsung donne vie à n'importe quelle photo ou portrait*, Le Big Data, 2019, https://www.lebigdata.fr/ia-samsung-photo-portrait

Utilisations inquiétantes

Lancée en Chine le 30 août 2019, l'application ZAO permet de remplacer un visage par un autre ("*face-swapping*") et notamment de remplacer en quelques secondes le visage des acteurs dans des vidéos de scènes connues par un autre à l'aide d'une simple photo[18]. Créée par l'entreprise chinoise MOMO Inc, initialement un éditeur d'application de rencontres, cette application connaît un immense succès. *"La capacité de créer des vidéos « deepfake » n'avait jusqu'ici jamais été proposée aussi facilement que dans l'interface de Zao. Il suffit à l'application d'une photo (ou une petite vidéo de son visage, afin d'obtenir un meilleur effet) pour remplacer le visage d'un acteur en quelques secondes. Et d'un clic pour partager le résultat, forcément amusant, sur les réseaux sociaux."*[19]

Ceci nous amène aux scandales qui affectent de nombreuses personnes, dont de nombreuses personnalités car leur visage est utilisé dans des vidéos pornographiques (notamment diffusées sur les plateformes Reddit et PornHub). Malgré l'annonce en février 2018 par les deux plateformes qu'elles condamnent les *deepfakes*, on trouve encore de multiples vidéos de "*deepfake porn*" sur leurs sites. Google de son côté interdit "*l'imagerie pornographique synthétique involontaire*" et permet aux victimes de *deepfake porn* de demander un blocage des vidéos dans lesquelles elles apparaissent. D'un point de vue légal néanmoins, bien qu'il soit possible d'attaquer quelqu'un qui vole une photo privée et la partage sur internet, peu de recours existent aujourd'hui aux Etats-Unis pour attaquer ceux qui partagent des contenus créés. Fin 2018, l'actrice Scarlett Johansson, victime de *deepfake porn*, dont une des vidéos cumule plus de 2 millions de vues, estime que se battre contre le *deepfake porn* est une cause perdue.

---

[18] L'application est exclusivement accessible depuis la Chine. *Can you believe your eyes? How deepfakes are coming for politics,*
https://www.google.com/url?q=https://www.ft.com/content/4bf4277c-f527-11e9-a79c-bc9acae3b654&sa=D&ust=1582556271698000&usg=AFQjCNFswch5I94ZUVyPrdEC2cVsvICDSg

[19]https://www.lemonde.fr/pixels/article/2019/09/03/a-peine-lancee-l-application-de-videos-deepfakes-zao-suscite-des-inquietudes_5505933_4408996.html

“

D'un point de vue légal néanmoins (…) peu de recours existent aujourd'hui aux Etats-Unis pour attaquer ceux qui partagent des contenus créés.

On peut noter néanmoins une avancée sur le sujet dans l'Etat de Virginie avec, en juillet 2019 l'interdiction du *deepfake porn*. Cette loi a ensuite été suivie par une loi en Californie qui permet à tous les californiens de porter plainte s'ils sont victimes d'un *deepfake* à caractère pornographique[20].

Au-delà du but premier d'amuser les populations, la chaîne Youtube *Facial Deception*, a publié une vidéo de Jim Carrey dans le corps de l'actrice américaine Alison Brie. Le but de cette vidéo est également de sensibiliser la population aux *fakes* et démontre que les *fakes* sont malheureusement de plus en plus faciles à réaliser pour ceux qui voudraient les utiliser à mauvais escient.

Source : Chaîne Youtube - Facial Deception[21]

[20] Kari Paul, *California makes 'deepfake' videos illegal, but law may be hard to enforce*, 2019, https://www.google.com/url?q=https://www.theguardian.com/us-news/2019/oct/07/california-makes-deepfake-videos-illegal-but-law-may-be-hard-to-enforce&sa=D&ust=1582556272252000&usg=AFQjCNFDv5BvCoRyuhSxcE0VWZIPDfEm1A

[21] Voir la vidéo ici: https://www.youtube.com/watch?v=b5AWhh6MYCg

“

Les fakes sont malheureusement de plus en plus faciles à réaliser pour ceux qui voudraient les utiliser à mauvais escient.

Les vidéos de Mark Zuckerberg ou de Jim Carrey avaient de leur côté avant tout été créées pour sensibiliser le public à la question de l'IA et plus spécifiquement des algorithmes de génération de contenu. Revenons tout d'abord sur la vidéo imitant Mark Zuckerberg. Elle a été créée par Bill Posters et Daniel Howe en collaboration avec le logiciel CannyAI[22], initialement dans le contexte d'une exposition artistique intitulée *Spectre*[23]. Bill Posters et Daniel Howe[24] sont deux "*hacktivistes*" britanniques qui se décrivent comme artistes chercheurs. CannyAI est une entreprise située en Israël qui produit un logiciel permettant de réaliser des doublages de vidéos dans de multiples langues, notamment la vidéo de Zuckerberg présentée précédemment. Cette vidéo fut postée par ses deux auteurs pour prendre Facebook à son propre piège, après le refus de l'entreprise de supprimer la vidéo de Nancy Pelosi. La vidéo de Zuckerberg a eu un impact important dans les nouvelles à la suite de quoi Facebook a annoncé, fin juin 2019, le lancement d'un projet en interne pour mieux contrôler les *deepfakes*. En janvier 2020, Facebook annonce l'interdiction des *deepfakes*.

Exemples Positifs

Les avancées dans le domaine de génération de vidéos sont impressionnantes et heureusement nous voyons apparaître de plus en plus d'exemple d'applications avec des buts positifs pour la société.

Nous notons notamment le lancement par Huawei d'une application proposant des livres avec vidéos en langue de signe pour aider les enfants à apprendre la langue des signes et leur faire

[22] Startup qui offre la possibilité de remplacer les dialogues d'une vidéo, et qui permet donc la création de *deepfakes*. CannyAI, https://www.cannyai.com/
[23] Bill Posters, *Spectre*, 2019, http://billposters.ch/projects/spectre/
[24] Daniel Howe, *Spectre*, 2019, http://rednoise.org/daniel/detail.html#spectre

découvrir le monde des livres : *"Ainsi, en 2018, en ayant recours à l'IA et à la réalité augmentée, nous avons créé StorySign, la première plateforme d'alphabétisation au monde à destination des enfants sourds. Apprendre à lire peut être difficile pour n'importe quel enfant. Cependant, pour les jeunes enfants atteints d'une surdité sévère ou profonde, cela peut représenter un défi de taille. Nous savions que notre technologie pourrait permettre à 32 millions d'enfants sourds à travers la planète d'accéder au royaume des livres et qu'elle pourrait les initier à un monde empreint d'imagination, de curiosité et de créativité. StorySign est une application mobile gratuite qui a pour but d'aider les enfants sourds à lire en traduisant des livres en langue des signes. À travers le pouvoir de l'IA et de la réalité augmentée, StorySign donne vie à ces livres et aide les enfants sourds à profiter du moment privilégié des histoires, comme cela devrait être permis à chaque enfant. "*[25]

1.4. Génération de texte

1.4.1. Open AI : GPT-2 et GPT-3

OpenAI est une entreprise à but lucratif plafonné (initialement une association à but non lucratif, mais les statuts ont changé en 2019) fondée en 2015, entre autres, par l'entrepreneur Elon Musk. Son objectif est de développer et promouvoir une intelligence artificielle humaine et positive pour la société.

En février 2019, OpenAI annonce avoir créé un algorithme de génération de langage bien trop puissant pour être divulgué au grand public : GPT-2 (*Generative Pretrained Transformer 2*) . Il fait suite à GPT (juin 2018). En effet, les chercheurs d'OpenAI estiment que GPT-2 est trop dangereux pour être publié en open source. L'algorithme permet en effet de reconstituer un texte à l'aide de quelques mots. Cette décision de garder leur code fermé au public fut largement critiquée par les communautés de l'IA. Après avoir choisi de publier seulement partiellement son code, OpenAI a été accusée de souhaiter faire le "buzz" alors que les résultats ne sont pas si impressionnants. Neuf

[25] https://consumer.huawei.com/fr/campaign/storysign/

mois après sa sortie, l'association a, finalement, publié entièrement le code de son algorithme en open source.

Nous mettons ci-dessous deux exemples de texte généré par GPT-2[26]:

```
Questions Q: What is love?
 A: love
 Q: How do you know if someone has a disease?
 A: look at them
 Q: When does the disease begin?
 A: when they are born
```

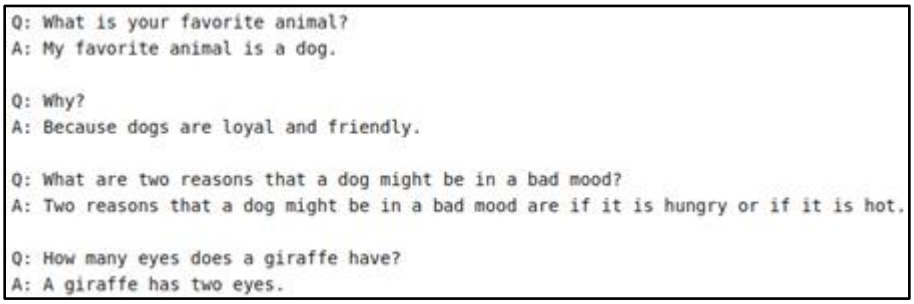
```
Links https://www.nytimes.com/2020/01/14/us/
trump-arrested-impeach.html
 Trump Is Arrested and Impeached

 WASHINGTON — President Trump, who has been accused of sexual assault
 by more than a dozen women, was arrested early Friday morning at his
 golf club in Bedminster, N.J., on charges that he had abused the
 power granted to him under an obscure 1799 law.

 The White House said Mr. Trump's arrest stemmed from "multiple
 allegations made against him," but it provided no further details.

 A lawyer for Mr. Trump, Rudolph W. Giuliani, told reporters outside
 the courthouse after the president left court that "the indictment is
 not based upon anything related to campaign activity."

 But several people close to Mr. Trump described him as furious about
 what they called a politically motivated prosecution.

 "The president's base is very upset," one person said.
```

Source : OpenAI

Comme nous le voyons dans l'exemple ci-dessus, l'algorithme peut servir à créer de faux articles réalistes. Il pourrait être utilisé à mauvais escient et notamment pour générer automatiquement des avis ou commentaires sur les réseaux sociaux, des spams, des faux articles scientifiques ou politiques, ou encore des fausses nouvelles.

[26] De nombreux autres peuvent être trouvés ici : https://openai.com/blog/better-language-models/#sample8

Environ un an plus tard, OpenAI produit une nouvelle version de son algorithme : GPT-3. Avec 175 milliards de paramètres cet algorithme successeur du GPT-2 est capable de générer non seulement du texte, mais aussi des lignes de codes, une page web, des tablatures de guitare ou bien même des valeurs manquantes dans des bases de données. Ayant comme but de révolutionner le monde de l'IA, GPT-3 constitue une nouvelle forme d'intelligence artificielle.



Source : Twitter[27]

En écrivant des questions aléatoires, GPT-3 est capable de répondre d'une façon assez logique et précise. Mais comment se traduit ce mécanisme ?

Appuyée sur du *deep learning*, GPT-3 cherche à comprendre le modèle dans les données et à trouver les *patterns* récurrents (i.e. un modèle simplifié d'une structure plus complexe). Ainsi, sans aucune intervention humaine, et en se basant sur des poids, il est capable de faire des prédictions à partir du texte fourni en entrée. Cet algorithme a été entraîné sur une énorme compilation de textes afin d'exploiter les régularités. Les nœuds du réseau de neurones stockent ultérieurement ces régularités en tant que connections pondérées[28].

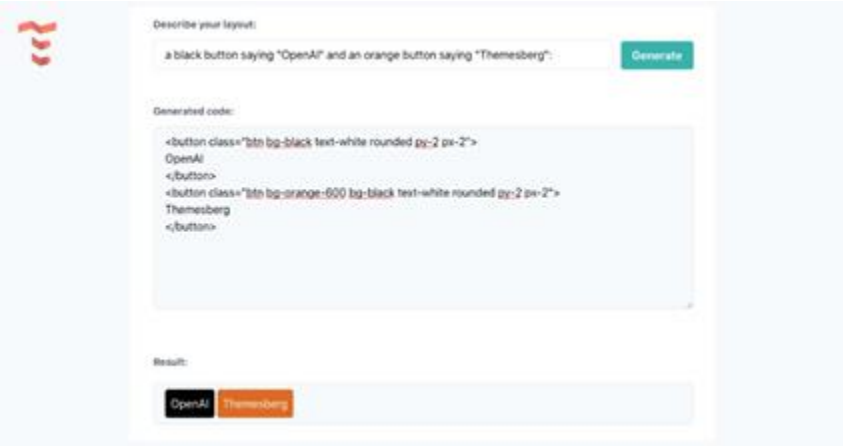
Pour illustrer très simplement son fonctionnement, prenons l'exemple de la phrase "Je voulais faire une omelette, alors je suis allé au frigo et j'en ai sorti des ___". Le blanc à la fin peut théoriquement être rempli avec n'importe quel mot. Mais le mot "œufs" a probablement un score assez élevé pour remplir ce vide dans la plupart des textes normaux. Une probabilité plus élevée

[27] A voir ici: https://twitter.com/lacker/status/1280276200582025216
[28] James Vincent, "OpenAI's latest breakthrough is astonishingly powerful, but still fighting its flaws", The Verge, 2020, https://www.theverge.com/

que, disons, le mot "éléphants". La probabilité de présence du mot “œufs” dans le texte est donc plus élevée que la probabilité de présence du mot “éléphants”.

> ***L'algorithme (…) pourrait servir à générer automatiquement des avis ou commentaires sur les réseaux sociaux, des spams, des faux articles scientifiques ou politiques, ou encore des fausses nouvelles.***

Grâce à sa capacité d’automatisation des tâches, GPT-3 pourra représenter dans le futur un atout important dans le monde du business, et surtout dans les entreprises de la *high tech* ou de la data. GPT-3 est notamment capable d’écrire un code HTML[29], comme le montre l’interface suivante :



Source: Zoltan Szogyenyi’s blog in the ThemesBerg, 2020

Malgré ces diverses caractéristiques, cela n’empêche pas l’existence de plusieurs inconvénients. Au niveau des entreprises, la complication essentielle réside dans le fait que GPT-3 nécessite beaucoup de données et de puissance de calcul. Son coût de compilation est important. Le texte

[29] Zoltan Szogyenyi, "We built an OpenAI powered Tailwind CSS code generator using GPT-3”, ThemesBerg, 2020, https://themesberg.com/

généré peut paraître impressionnant, mais l'algorithme commet davantage d'erreurs quand il s'agit de demandes plus complexes.

Q: When counting, what number comes before 1000?
A: Nine hundred and ninety-nine comes before 1000.

Q: When counting, what number comes before 10,000?
A: Nine thousand and ninety-nine comes before 10,000.

Q: When counting, what number comes before a million?
A: Nine hundred thousand and ninety-nine comes before a million.

Source : Mind Matters, 2020[30]

Ainsi, dans l'exemple ci-dessus, GPT-3 est incapable, à partir d'un certain nombre de chiffres, d'être performant dans les séries arithmétiques tout en se basant sur des données entrainées. Par conséquent, il s'avère que GPT-3 peut faire des erreurs que le cerveau humain ne ferait pas.

1.4.2. Génération de partitions

Dans le domaine musical nous pouvons aussi observer des applications dans la génération de musique. Les réseaux récurrents avec mémoire temporelle ont fait leurs preuves : la création d'un morceau de jazz est devenue un exercice classique de code avec l'algorithme LSTM, grâce au cours de Andrew Ng de l'université de Stanford[31]. En 2017, le modèle MidiNet[32], utilisant des *GANs* avec des CNNs (*Convolutional Neural Network*)[33] est dédié à la génération de musique.

Dans la même veine, Huawei diffuse en 2017 une version achevée de la huitième symphonie de Schubert, débutée en 1822 et connue pour n'avoir jamais été terminée. "*Grâce à la puissance de l'IA, Huawei a appris au Mate 20 Pro à composer et ainsi créer les troisièmes et quatrièmes*

[30] A voir ici: https://mindmatters.ai/2020/07/gpt-3-is-mindblowing-if-you-dont-question-it-too-closely/
[31] deeplearning.a, *Recurrent Neural Networks - Why sequence models?*, https://cs230.stanford.edu/files/C5M1.pdf
[32] Li-Chia Yang, Szu-Yu Chou, Yi-Hsuan Yang, *MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation*, 2017, https://arxiv.org/abs/1703.10847
[33] En apprentissage automatique, un réseau de neurones convolutifs ou réseau de neurones à convolution (en anglais CNN ou ConvNet pour Convolutional Neural Networks) est un type de réseau de neurones artificiels acycliques (*feed-forward*), dans lequel le motif de connexion entre les neurones est inspiré par le cortex visuel des animaux.

mouvements de la célèbre « symphonie inachevée » de Schubert. Cela a été rendu possible en utilisant la puissance et l'intelligence de ce smartphone exceptionnel conçu pour effectuer des tâches faisant appel à l'IA. Le Mate 20 Pro a d'abord écouté les deux premiers mouvements de la symphonie n° 8 de Schubert, puis a analysé les éléments qui la rendent si particulière. En se basant sur cette analyse, le smartphone a ensuite pu générer les deux mouvements suivants."[34] Même si les critiques ne sont pas tous convaincus de retrouver le style de Schubert dans les notes générées, il n'en demeure pas moins que c'était une action de marketing réussie de la part de Huawei pour mettre en avant ses avancées sur les sujets IA.

Huawei diffuse en 2017 une version achevée de la huitième symphonie de Schubert (...). Même si les critiques ne sont pas tous convaincus de retrouver le style de Schubert dans les notes générées, il n'en demeure pas moins que c'était une action de marketing réussie de la part de Huawei pour mettre en avant ses avancées sur les sujets IA.

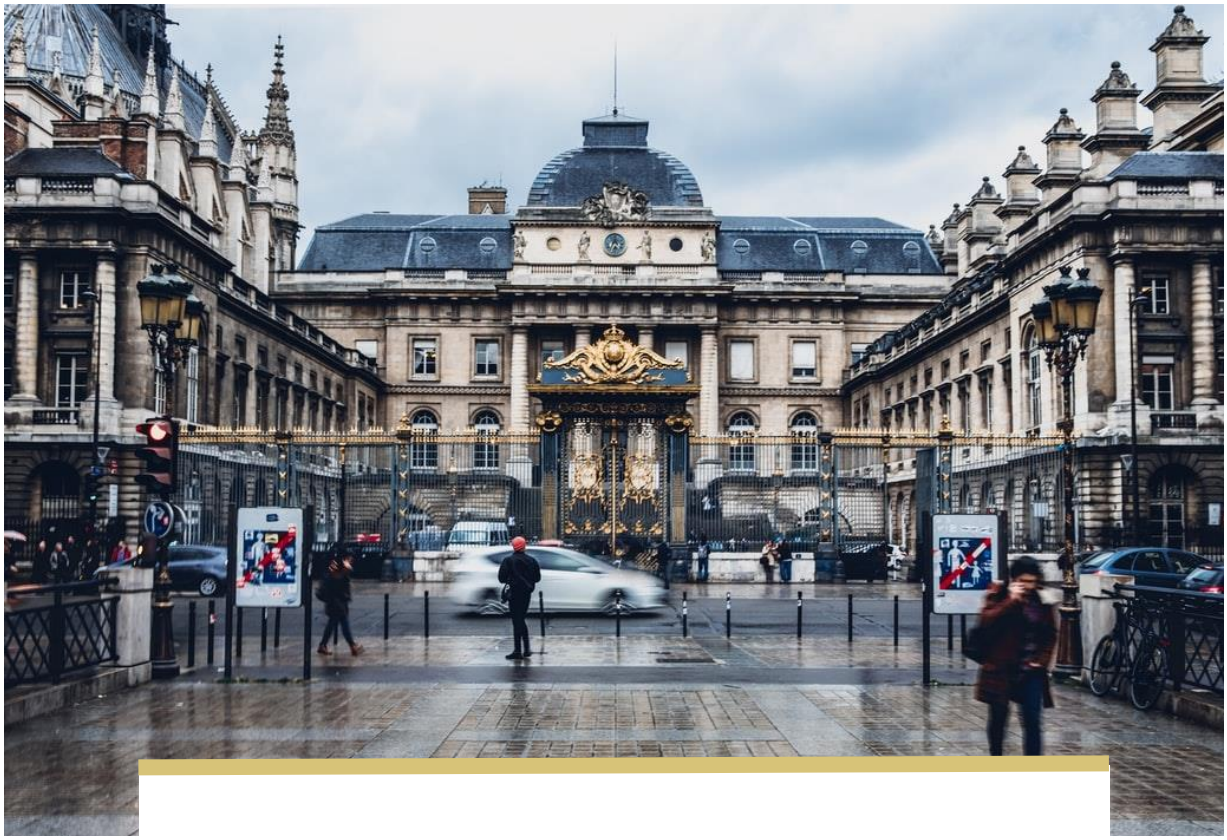
1.5. Génération de données

Les algorithmes de génération de contenu s'avèrent de plus en plus utiles pour compléter des données manquantes. En effet, par sa capacité à comprendre un ensemble de données et à en générer d'autres, ces algorithmes peuvent être utilisés pour créer des données réalistes. Ceci est utile pour de nombreux secteurs, par exemple en assurance pour estimer le risque de catastrophe

[34] https://consumer.huawei.com/fr/campaign/unfinishedsymphony/

naturelle, sur le risque de pandémie ou pour tout événement pour lequel nous manquons de données historiques.

Les algorithmes de génération de contenu s'avèrent de plus en plus utiles pour compléter des données manquantes (...) Ceci est utile pour de nombreux secteurs, par exemple en assurance pour estimer le risque de catastrophe naturelle, sur le risque de pandémie ou pour tout événement pour lequel nous manquons de données historiques.



2. Les enjeux éthiques liés aux algorithmes de génération de contenu

2.1. Contexte

"Chaque bond en avant du progrès technique laisse le développement intellectuel relatif des masses d'un pas en arrière, et cause donc une chute du thermomètre de la maturité politique. Il faut parfois des dizaines d'années, parfois des générations, pour que le niveau de compréhension d'un peuple s'adapte graduellement au nouvel état des choses…"

Arthur Koestler, Le Zéro et L'Infini.

Les algorithmes de génération de contenu posent de nombreuses questions qui portent sur le respect de la liberté individuelle, le respect d'autrui et plus généralement sur la cohabitation dans une société démocratique.

Le développement et l'utilisation de programmes qui peuvent générer des contenus *deepfakes* devraient conduire à nous (re)poser la question : est-ce que tout ce qui est techniquement possible doit être entrepris ? Qu'est-ce qui doit être régulé ?

2.1.1. Historique

La diffusion de fausses informations ou de faux documents n'est pas chose nouvelle[35]. Déjà le Pharaon Ramsès II avait fait croire dans des inscriptions que les Egyptiens avaient vaincu les Hittites lors de la bataille de Kadesh alors qu'en réalité, il n'y avait pas eu de vainqueur. L'un des premiers faux documents utilisés pour justifier une action est la *Donation de Constantin*, lettre par laquelle l'empereur Constantin I[er] transférait en l'an 315 tout son pouvoir sur l'Occident au pape Sylvestre, l'empereur Constantin devenant uniquement alors empereur d'Orient. La donation est évoquée pour la première fois en 979, et sa fausseté ne sera prouvée que 500 années plus tard.

Les Protocoles des Sages de Sion quant à eux, textes publiés en Russie au début du XX[e], dévoilent les prétendues intentions des communautés juives et franc-maçonnes de dominer le monde. Ces

[35] Wikipedia, *Fake news*, https://en.wikipedia.org/wiki/Fake_news#Ancient

textes ont malheureusement été mis en avant par Hitler, notamment dans *Mein Kampf* pour justifier la théorie du complot juif. Ils sont encore diffusés dans les milieux antisémites.

Côté manipulation photographique on peut noter les photos des dignitaires soviétiques qui étaient modifiées pour faire disparaître les dirigeants tombés en disgrâce. Ainsi, cette photo datant des années 30[36] :

Photos de Staline avec et sans Nikola Yezhov, ancien bras droit de Staline.

Source : Wikimedia, 1930

Quelques dizaines d'années auparavant et dans un autre contexte, on peut mentionner cette photo du général Ulysse Grant datant de la guerre civile américaine qui est en fait une composition de trois photos différentes[37].

[36] Wikimedia, *URSS, 1930,* https://curiummag.com/wp-content/uploads/2016/08/staline-1024x462.png
[37] Wikpedia, *Photo manipulation*, https://en.wikipedia.org/wiki/Photo_manipulation

Photo du général Ulysse Grant

Source : Wikimedia, 1930

La manipulation des textes et des images à des fins politiques n'est donc pas une nouveauté.

2.1.2. Les algorithmes de génération de contenu

Les algorithmes de génération de contenu sont utilisés pour créer des contenus qui peuvent être ludiques, drôles, et ne sont bien sûr pas appliqués uniquement pour créer des *deepfakes*.

En plus de création d'image, de son, de vidéos, ils peuvent être consacrés à compléter un ensemble à faible nombre de données. Ceci peut s'avérer très utile pour améliorer les modèles qui apprennent sur peu de données, par exemple prédire l'impact d'une épidémie, ou évaluer un risque de cybersécurité pour mieux estimer le prix d'une assurance de cybersécurité.

Dans le domaine de l'imagerie médicale, où le nombre d'exemples disponible est souvent limité, ils permettent de créer de nouvelles images qui peuvent servir à entraîner des modèles de *deep learning*[38].

Dans le domaine artistique, du *design* ou de la mode, les *GANs* permettent de créer des *designs* nouveaux ainsi que des *designs* personnalisés et uniques. Ils peuvent aussi trouver de nombreuses applications dans le domaine des jeux vidéo[39].

Les *GANs* permettent de créer des environnements fictifs utiles qui peuvent s'avérer très opportuns pour former des machines voire des humains à de nouvelles tâches.

Le PDG de Ekimetrics, Jean-Baptiste Bouzige, considère que les *GANs* permettraient de réduire les coûts de construction en facilitant la modélisation de bâtiment très en amont de leur *design*, par exemple pour assurer une haute efficacité énergétique ou autre contrainte technique contenant beaucoup de paramètres.

2.1.3. Démocratisation de la production des contenus

Une autre différence majeure avec les expériences historiques est que chaque individu peut être tour à tour producteur, consommateur et relayeur de contenus et bien entendu citoyen.

Pour prendre l'exemple du journalisme, bien qu'il soit délicat pour un journaliste d'être impartial et qu'il soit impossible de communiquer une information sans biais, le métier de journaliste s'est organisé progressivement et une déontologie du traitement de l'information s'est structurée, autour notamment de la charte de Munich[40]. Le journaliste s'engage notamment à vérifier les faits

[38] Maayan Frid-Adar et a., "Synthetic Data Augmentation using GAN for Improved Liver Lesion Classification ", Department of Biomedical Engineering, Tel Aviv University, 2018

[39] Wang-Cheng Kang et al., "Visually-Aware Fashion Recommendation and Design with Generative Image Models" , 2017

[40] Charte de Munich : Déclaration des droits et des devoirs des journalistes de 1971 et Charte du Syndicat National des journalistes de 1918, révisée en 1938

qu'il relate. Aujourd'hui, cette déontologie n'est pas partagée par tous les producteurs de contenus et de nombreuses *fake news*[41] circulent sur le web.

Il est impossible de forcer les plateformes à assurer l'authenticité de tout contenu qu'elle produit car, nous l'avons vu, il est parfois impossible d'identifier le vrai du faux. Un faux contenu n'est par ailleurs pas nécessairement malveillant : une caricature ne vise par exemple pas nécessairement à nuire. Néanmoins, une mesure simple nous semble pertinente pour mieux encadrer le faux contenu : une obligation pour les auteurs de déclarer un contenu transformé (au-delà du simple filtre et des corrections cosmétiques). Ceci permettrait de responsabiliser les auteurs de faux contenus et aussi de mieux protéger les utilisateurs de plateformes.

2.1.4. Les réseaux sociaux : diffusion accélérée, biais de vérité en mettant en avant les faux contenus

Il est indéniable qu'aujourd'hui un contenu peut se propager de manière quasi-instantanée, et que le nombre de personnes pouvant être atteintes a explosé. Selon une étude du quotidien américain *International Business Times,* 60% des *millenials* américains utilisent les réseaux sociaux comme Twitter et Facebook pour se tenir au courant de l'actualité plutôt que les médias traditionnels. Par conséquent, ils sont beaucoup plus exposés à un contenu viral, qui est mis en avant par les plateformes, selon des lois de viralité et de nombre de *likes*.

En effet, pour des raisons économiques d'optimisation du nombre de vues (et de profit), les plateformes ont intérêt à favoriser la distribution (ou viralité) de vidéos attirant le regard (ou la polémique) qui génèrent de multiples transmissions d'individu à individu.

Nous notons par exemple le phénomène des *mèmes*[42] sur internet : excitants, ludiques, ils se répandent rapidement sur le web et se révèlent être un excellent outil de communication, ce

[41] Les fakes news représentent des informations qui sont fausses largement diffusées sur les médias

[42] Un mème est un "concept (texte, image, vidéo) massivement repris, décliné et détourné sur Internet de manière souvent parodique, qui se répand très vite, créant ainsi le buzz." Larousse, https://www.larousse.fr/dictionnaires/francais/m%C3%A8me/10910896

qu'ont bien compris de nombreux publicitaires, mais aussi de nombreuses personnes cherchant à défendre leur cause politique. Le procédé n'est pas en soi si différent des caricatures utilisées depuis des centaines d'années. La différence avec le passé n'est pas tant sur l'attrait du scandale, l'utilisation du rire, de la parodie, pour décrédibiliser ou faire passer un message, mais plutôt sur d'une part cette vitesse décuplée dans la diffusion qui empêche toute réaction efficace et d'autre part sur ce biais qui fait que ce contenu prend plus de place qu'un contenu banal qui tente d'informer.

Lorsqu'un *deepfake* avec le Président Obama est mis en ligne, même s'il est faux et que nombreux confirment qu'il est faux, il est très compliqué d'en limiter son impact.

L'ancienne ministre Nathalie Kosciusko-Morizet[43] que nous avons rencontrée met en avant l'existence de deux risques majeurs : la rapidité de diffusion de l'information et la viralité. Pour le premier, cela représente plus précisément le décalage qui existe entre la diffusion des fausses nouvelles et celles des démentis. En ce qui concerne le second, il est aggravé par le déclin des intermédiaires suite à l'expansion du numérique, où, selon Nathalie Kosciusko-Morizet :

"*Le premier risque est la rapidité de la diffusion des informations. Précisément, le décalage entre la vitesse de diffusion des mensonges et celle des démentis. Il existe des études documentées sur ce sujet. Le deuxième risque est celui de la viralité. Avec le numérique vient le déclin des intermédiaires, qui servaient de modérateurs. C'est la prime à l'émotion, à la radicalité, à l'excès. L'appétence pour le scandale et la polémique joue à plein, d'autant que cette propension naturelle de l'âme humaine est favorisée par les algorithmes qui font remonter en haut des pages les opinions les plus controversées. Le problème est accru par la rémanence des informations sur le net. Même démenties ou dépassées, les informations restent en ligne.*"

Une étude publiée en 2018 confirme les propos de l'ancienne ministre, en affirmant "*qu'une fake news se répand 6 fois plus vite sur Twitter qu'une information issue de fact-checking.*"[44]

[43] Nathalie Kosciusko-Morizet, *Cyberespace et déstabilisation de la démocratie*, 2020 https://www.linkedin.com/pulse/cyberespace-et-d%C3%A9stabilisation-de-la-d%C3%A9mocratie-kosciusko-morizet/
[44] Vosoughi, S., Roy, D., & Aral, S. (2018). *The spread of true and false news online*. Science, 359(6380), 1146-1151.

2.2. Qu'est-ce que l'éthique ?

Le mot "éthique" fait partie de ces tiroirs qui sont pratiques pour déplacer un débat tout en l'élevant, sans qu'on sache parfois bien ce que le mot signifie. Sans entrer dans de longs développements, il nous semble important de rappeler sa signification et de voir comment il a pu trouver un sens concret et pratique dans le cas de la médecine.

Dans un texte de 1990, le philosophe Paul Ricœur définissait l'éthique comme étant " *la visée d'une vie accomplie sous le signe des actions estimées bonnes*."[45] Le philosophe, Dominique Lecourt, définit l'éthique comme "*une réflexion, un questionnement qui porte sur des dilemmes. Face à des situations complexes, il y a un choix à faire entre plusieurs réponses qui sont toutes insatisfaisantes.*"[46] L'éthique est donc une réflexion individuelle sur ce qui semble bon et doit être choisi, même si aucune contrainte extérieure (légale ou morale) n'impose ce choix.

Cette notion de "bien" dans une société démocratique n'est pas facile à définir, mais on sent qu'elle peut facilement entrer en contradiction avec la morale des affaires, qui pose comme limite de faire tout ce qui est légalement autorisé pour optimiser la performance économique. Ainsi, un diffuseur peut trouver des justifications économiques de laisser circuler un *deepfake* alors même qu'il peut estimer à titre individuel que ce n'est pas éthiquement correct.

2.3. Premières actions pour une éthique dans le net

Depuis quelques années, plusieurs initiatives ont émergé tant des acteurs du net que des autorités publiques pour élaborer des chartes de l'éthique.

Côté entreprises, nous pouvons citer notamment *Partnership on AI*[47] dont font partie les géants Amazon, Facebook, Google, Microsoft, IBM, Apple mais aussi le géant chinois Baidu depuis 2018.

[45] Paul Ricœur, *Éthique et morale*, 1990
[46] Simon SUTOUR et Jean-Louis LORRAIN au nom de la commission des affaires européennes, Rapport au Sénat, 2013
[47] The Partnership on AI, https://www.partnershiponai.org/

Cette coalition, fondée en 2016, a pour but d'éduquer le public sur l'IA mais aussi de définir des bonnes pratiques en matière d'intelligence artificielle.

De surcroît, d'autres acteurs ont également proposé des principes éthiques à suivre quand on fait de l'algorithmique.

Commission Européenne

La Commission Européenne a récemment publié un livre blanc sur l'intelligence artificielle[48] qui reprend entre autres les conclusions de son rapport de juin 2018. Selon elle, "*l'IA peut avoir de nombreux effets positifs, notamment en matière de renforcement de la sécurité des produits et des procédés, mais aussi des effets négatifs. Elle peut être préjudiciable tant sur le plan matériel (en matière de sécurité et de santé des personnes : pertes humaines, dommages aux biens) qu'immatériel (atteinte à la vie privée, restrictions du droit à la liberté d'expression, dignité humaine, discrimination à l'embauche par exemple), et impliquer des risques de types très divers*."

Les principes éthiques avancés par la Commission Européenne sont :

1. **Action humaine et contrôle humain**
2. **Robustesse, technique et sécurité**
3. **Respect de la vie privée et gouvernance des données**
4. **Transparence**
5. **Diversité, non-discrimination et équité**
6. **Bien-être sociétal et environnemental**
7. **Responsabilité**

[48] Commission Européenne, *Une approche européenne axée sur l'excellence et la confiance*, 2020, https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_fr.pdf

Source : European Commission, *Ethics Guidelines for Trustworthy AI*

L'UE propose également de veiller à ce qu'une formation dans le domaine de l'éthique en matière d'IA soit systématiquement dispensée à une nouvelle génération d'experts et d'adopter une évaluation pour une IA digne de confiance lors de la mise au point, du déploiement ou de l'utilisation de systèmes d'IA, et l'adapter au cas d'utilisation spécifique du système.

OCDE

L'OCDE a également pris part au débat autour de l'intelligence artificielle en publiant une liste de recommandations et de de principes qui s'avèrent très similaires[49] à ceux énoncés par la Commission Européenne :

- **Croissance inclusive, développement durable et bien-être :**

  Les parties prenantes devraient s'engager de manière proactive dans une gestion responsable de l'IA fiable dans la recherche de résultats bénéfiques pour les personnes et

[49] OECD Principles on AI, https://www.oecd.org/going-digital/ai/principles/

la planète, telles que l'augmentation des capacités humaines et l'amélioration de la créativité, la promotion de l'inclusion des populations sous-représentées, la réduction des inégalités économiques, sociales, de genre et autres, et la protection des environnements naturels, dynamisant ainsi la croissance inclusive, le développement durable et le bien-être.

- **Valeurs et équité centrées sur l'homme :**

Les acteurs de l'IA devraient respecter l'état de droit, les droits de l'homme et les valeurs démocratiques tout au long du cycle de vie du système d'IA. Ceux-ci incluent la liberté, la dignité et l'autonomie, la vie privée et la protection des données, la non-discrimination et l'égalité, la diversité, l'équité, la justice sociale et les droits du travail internationalement reconnus.

À cette fin, les acteurs de l'IA devraient mettre en œuvre des mécanismes et des garanties, tels que la capacité de détermination humaine, adaptés au contexte et conformes à l'état de la technique.

- **Transparence et explicabilité :**

Les acteurs de l'IA devraient s'engager à assurer la transparence et la divulgation responsable des systèmes d'IA. À cette fin, ils devraient fournir des informations significatives, adaptées au contexte et conformes à l'état de la technique :

  - Pour favoriser une compréhension générale des systèmes d'IA.
  - À sensibiliser les parties prenantes à leurs interactions avec les systèmes d'IA, y compris sur le lieu de travail.
  - Permettre aux personnes affectées par un système d'IA de comprendre le résultat, etc.
  - Permettre à ceux qui sont affectés par un système d'IA de contester ses résultats sur la base d'informations claires et faciles à comprendre sur les facteurs et la logique qui a servi de base à la prédiction, à la recommandation ou à la décision.

- **Robustesse, sécurité et sûreté :**

  Les systèmes d'IA doivent être robustes, sécurisés et sûrs tout au long de leur cycle de vie afin que, dans des conditions d'utilisation normale, prévisible ou abusive, ou dans d'autres conditions défavorables, ils fonctionnent de manière appropriée et ne présentent pas de risque de sécurité déraisonnable.

  À cette fin, les acteurs de l'IA devraient garantir la traçabilité, y compris en ce qui concerne les ensembles de données, les processus et les décisions prises au cours du cycle de vie du système d'IA, pour permettre l'analyse des résultats du système d'IA et des réponses à l'enquête, appropriée au contexte et cohérente avec l'État d'art.

  Les acteurs de l'IA devraient, en fonction de leurs rôles, du contexte et de leur capacité à agir, appliquer une approche systématique de gestion des risques à chaque phase du cycle de vie du système d'IA sur une base continue pour faire face aux risques liés aux systèmes d'IA, y compris la confidentialité, le numérique sécurité, sûreté et parti pris.

- **Responsabilité :**

  Les acteurs de l'IA devraient être responsables du bon fonctionnement des systèmes d'IA et du respect des principes ci-dessus, en fonction de leurs rôles, du contexte et conformément à l'état de l'art.

Rapport CNIL

La CNIL, Commission nationale de l'informatique et des libertés, a publié un rapport en 2017 sur les enjeux éthiques des algorithmes et de l'intelligence artificielles[50] qui définissait deux principes fondateurs *"pour une intelligence artificielle au service de l'homme"* : un principe de loyauté et un principe de vigilance (cf. Encadré).

[50] CNIL, *How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence*, 2017, https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf

Rapport CNIL 2017

Vers l'affirmation de deux principes fondateurs : loyauté et vigilance

Un principe de *loyauté* appliqué à tous les algorithmes et intégrant les impacts collectifs, et pas seulement personnels, de ces derniers. Tout algorithme, qu'il traite ou non des données personnelles, doit être loyal envers ses utilisateurs, non pas seulement en tant que consommateurs, mais également en tant que citoyens, voire envers des communautés ou de grands intérêts collectifs dont l'existence pourrait être directement affectée. **L'intérêt des utilisateurs doit primer**. Par exemple, un tel principe pourrait avoir vocation à s'appliquer à l'impact potentiel des réseaux sociaux sur la structure du débat public dans nos démocraties (segmentation du corps politique par le ciblage de l'information) ou à celui d'algorithmes de police prédictive sur des communautés ou quartiers entiers.

Un principe de *vigilance/réflexivité* : il s'agit d'organiser une forme de questionnement régulier, méthodique et délibératif à l'égard de ces objets mouvants. Ce principe constitue une réponse directe aux exigences qu'imposent ces objets technologiques du fait de leur nature imprévisible (inhérente au *machine learning*), du caractère très compartimenté des chaînes algorithmiques au sein desquels ils s'insèrent et, enfin, de la confiance excessive à laquelle ils donnent souvent lieu. C'est l'ensemble des maillons de la chaîne algorithmique (concepteurs, entreprises, citoyens) qui doivent être mobilisés pour donner corps à ce principe, au moyen de procédures concrètes (par exemple, des comités d'éthique assurant un dialogue systématique et continu entre les différentes parties-prenantes).

Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires

Un groupe de travail de la Commission européenne "pour l'efficacité de la justice", a publié une charte éthique européenne sur l'utilisation de l'IA dans les systèmes juridiques en décembre 2018[51]. Cette charte met en avant cinq principes clés afin de lutter contre les dérives de l'intelligence artificielle et pour fixer un cadre éthique à l'utilisation de l'intelligence artificielle :

- **Principe de respect des droits fondamentaux :** visant à assurer une conception et une mise en œuvre des outils et des services d'intelligence artificielle qui soient compatibles avec les droits fondamentaux.
- **Principe de non-discrimination :** visant à prévenir spécifiquement la création ou le renforcement de discriminations entre individus ou groupes d'individus.
- **Principe de qualité et sécurité :** visant, en ce qui concerne le traitement des décisions juridictionnelles et des données judiciaires, à utiliser des sources certifiées et des données intangibles avec des modèles conçus d'une manière multidisciplinaire, dans un environnement technologique sécurisé.
- **Principe de transparence, de neutralité et d'intégrité intellectuelle :** visant à rendre accessibles et compréhensibles les méthodologies de traitement des données et à autoriser les audits externes.
- **Principe de maîtrise par l'utilisateur :** visant à bannir une approche prescriptive et permettre à l'usager d'être un acteur éclairé et maître de ses choix.

Elle conclut son rapport en expliquant que les évolutions technologiques doivent être encadrées par une "cyberéthique" qui "*doit s'accompagner d'une formation à large échelle des acteurs, depuis les concepteurs des algorithmes et les legaltechs jusqu'à leurs utilisateurs. De nouvelles humanités,*

[51]Commission Européenne pour l'efficacité de la Justice (CEPEJ), Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires et leur environnement, 2018 https://www.apda.ad/sites/default/files/2018-12/Charte%20ethique%20FR%20pour%20publication%204%20d%C3%A9cembre%202018.docx.pdf

transdisciplinaires, devraient être partagées entre tous pour que l'IA devienne le vecteur de développements positifs pour l'humanité et renforce les garanties de la démocratie et de l'État de droit."

2.4. Où en sont les plateformes internet ?

Les plateformes internet, notamment les réseaux sociaux, jouent un rôle majeur dans la diffusion de l'information, et par ricochet dans les contenus *deepfakes*. En effet, que ce soit en France[52], au Royaume-Uni [53] ou aux Etats-Unis, des études ont montré que la population s'informe davantage de l'actualité via internet et les réseaux que via les canaux traditionnels comme la radio ou la presse écrite. Nous avons déjà mentionné l'étude du quotidien américain *International Business Times*. Chez les Français de moins de 34 ans, les réseaux sociaux sont même la première source d'information, devant la télévision. Or, par leur fonctionnement et la personnalisation des contenus qu'ils diffusent, les réseaux sociaux peuvent favoriser la polarisation des opinions. En effet, individus et groupes polarisés s'intéressant à des informations erronées ou manipulées — qu'ils le sachent ou non — tendent à s'exposer à davantage de contenus du même type. Ce faisant, les réseaux sociaux peuvent renforcent les croyances et contribuer à polariser davantage.

Pourtant, si les géants du net ont conscience des enjeux et risques éthiques posés par les *fake news*, leur réponse à ce problème ne semble pas unanime. L'adoption universelle d'une charte éthique ne semble pas à l'ordre du jour tant la position des acteurs diffère. Chaque plateforme prend donc ses propres mesures.

---

[52] *Noémie Bonnin, Les réseaux sociaux première source d'info en ligne chez les personnes sensibles aux théories du complot, 2019,* https://www.francetvinfo.fr/internet/reseaux-sociaux/info-franceinfo-les-reseaux-sociaux-premiere-source-d-info-en-ligne-chez-les-personnes-sensibles-aux-theories-du-complot_3191963.html

[53] *CCFI, UK : la moitié de la population s'informe par les réseaux sociaux, 2019,* https://www.ccfi.asso.fr/uk-la-moitie-de-la-population-sinforme-par-les-reseaux-sociaux/

2.4.1. Facebook

En mai 2019, Facebook est sous le feu des critiques, le réseau social a laissé circuler une vidéo déformée de Nancy Pelosi, politicienne démocrate et présidente de la chambre des représentants des États-Unis, dans laquelle elle semble ivre. Devenue virale, la vidéo a été vue par plus de 2 millions de personnes en deux jours et a permis aux équipes de Donald Trump de la décrédibiliser. Facebook ne s'estimant ni garant, ni juge de la véracité des vidéos, prône la liberté d'expression et annonce qu'il ne supprimera pas de *deepfakes*. En juin 2019, Mark Zuckerberg, fondateur et Directeur Général de Facebook est à son tour mis en scène dans une vidéo déformée. Son personnage y annonce qu'il est prêt à se servir des milliards de données volées aux citoyens pour contrôler le futur et qu'il doit tout à "Spectre."

Néanmoins en septembre 2019, Facebook, en partenariat avec Amazon et Microsoft, lance un concours de détection des *deepfakes*, le *Deepfake Detection Challenge* (DFDC), pour développer des algorithmes capables d'identifier les vidéos et images *deepfakes*[54].

Depuis, Facebook a mis en place une politique de lutte contre les *deepfakes*. Depuis janvier 2020, celles-ci seront désormais supprimées de la plateforme[55] :

- Si le document "*a été édité ou synthétisé, au-delà des ajustements pour la clarté ou la qualité, d'une manière qui n'est pas évidente pour un individu moyen et qui pourrait induire en erreur en laissant penser qu'un sujet de la vidéo a dit des mots qu'il n'a pas réellement dit.*"
- Si "*c'est le produit de l'intelligence artificielle ou de l'apprentissage automatique qui fusionne, remplace ou superpose le contenu sur une vidéo, donnant ainsi l'impression qu'elle est authentique*."

[54] Facebook AI, *Deepfake Detection Challenge Dataset*, 2020, https://deepfakedetectionchallenge.ai/
[55] Facebook, *Enforcing Against Manipulated Media*, 2020, https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/

Bien évidemment, cette politique "*ne s'étend pas aux contenus parodiques ou satiriques, ni aux vidéos qui ont été modifiées uniquement pour omettre ou changer l'ordre des mots*." Les caricatures et l'usage des *deepfakes* pour des parodies sont autorisées.

Facebook ne demande néanmoins pas aux auteurs plus de transparence pour prévenir les utilisateurs de l'usage d'algorithmes de génération de contenu.

2.4.2. Twitter

Twitter a annoncé en février une nouvelle politique de lutte contre les *deepfakes*[56]. Le réseau social signalera les comptes qui partagent des documents *deepfakes*, avertira les comptes utilisateurs qui souhaitent partager ces documents ("*retweet*") et fournira les raisons pour lesquelles le document peut être un *deepfake*.

Source : Twitter

D'abord, Twitter mentionne que pour identifier un faux contenu, les facteurs suivants seront pris en compte :

[56] Yoel Roth & Ashita Achuthan, *Building rules in public: Our approach to synthetic & manipulated media*, 2020, https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html

- Si le contenu a été considérablement modifié d'une manière qui altère fondamentalement sa composition, sa séquence, son rythme ou son cadrage ;
- Toute information visuelle ou auditive (comme de nouvelles images vidéo, des données audios doublées ou des sous-titres modifiés) ayant été ajoutée ou supprimée ;
- Si les médias représentant une personne réelle ont été fabriqués ou simulés.

Si un contenu répond aux conditions citées ci-dessus alors Twitter pourra prendre diverses mesures, reprises dans le tableau ci-dessous, afin de limiter la portée et l'impact négatif du faux contenu. La cas échéant, Twitter pourra fournir une information certifiée sur le sujet à l'image de que la plateforme a fait sur certains Tweets de Donald Trump dénonçant le vote par correspondance[57].

Les mesures suivantes seront mises en place dans les cas précédemment cités (en fonction de la situation) :

- Appliquer un label sur le Tweet.
- Montrer un avertissement aux utilisateurs avant de retweeter ou aimer un Tweet.
- Réduire la visibilité du Tweet sur le réseau et/ou empêcher qu'il soit recommandé.
- Fournir des explications ou des précisions supplémentaires, le cas échéant, comme une page d'accueil avec plus de contexte ou un lien vers une source certifiée.

Is the media significantly and deceptively altered or fabricated?	Is the media shared in a deceptive manner?	Is the content likely to impact public safety or cause serious harm?	
✓	✗	✗	Content **may** be labeled
✓	✗	✓	Content is **likely** to be labeled, or **may** be removed.
✓	✓	✗	Content is **likely** to be labeled.
✓	✓	✓	Content is **very likely** to be removed.

Source : Twitter

[57]https://www.franceinter.fr/monde/pour-la-premiere-fois-un-message-de-donald-trump-signale-comme-mensonger-par-twitter

2.4.3. Google & Youtube

En tant que première plateforme vidéo au monde, Youtube joue nécessairement un rôle majeur dans la diffusion des vidéos *deepfakes*. Youtube a été le premier grand réseau à avoir retiré la vidéo trafiquée de Nancy Pelosi. Les responsables de Youtube déclarent vouloir "*devenir une source d'information plus fiable et de "promouvoir un discours politique sain"*[58]. Le réseau annonçait alors vouloir "*supprimer les contenus manipulés techniquement ou falsifiés dans le but de tromper les utilisateurs et qui posent un risque flagrant*". Une position qui se rapproche de celle des autres grandes plateformes.

Plus largement, Google (qui détient Youtube) a pour politique de ne pas autoriser des documents trompeurs ("*deceptive practice*"), comme ils l'affirment dans une réponse à la *House Permanent Select Committee on Intelligence* (une commission permanente du Congrès des États-Unis sous la responsabilité de la Chambre des représentants)[59] [60]. De plus ils collaborent avec le FBI afin d'identifier les utilisateurs malveillants et de désactiver leurs comptes.

[58]Le Figaro, *Deepfake, vidéos falsifiées... Quelle politique appliquent les différents réseaux sociaux?*, 2020 https://www.lefigaro.fr/secteur/high-tech/deepfake-videos-falsifiees-quelle-politique-appliquent-les-differents-reseaux-sociaux-20200206

[59] KaranBhatia, 2020, https://schiff.house.gov/imo/media/doc/KB%20Schiff%20response%207-2019%20(1).pdf

[60] *"Nous prenons la menace des médias manipulés très au sérieux, qu'il s'agisse de modifications générées par l'IA ou de faible technologie. YouTube a des politiques claires qui décrivent quel contenu n'est pas acceptable à publier et nous supprimons les vidéos violant ces politiques lorsqu'elles nous sont signalées. Nous nous efforçons toujours d'investir et d'améliorer nos processus et notre technologie pour appliquer nos directives, y compris contre les menaces potentielles liées aux supports synthétiques et à la désinformation. Le règlement de la communauté YouTube interdit certaines pratiques trompeuses visant à tirer parti de la communauté YouTube, y compris dans certains contextes impliquant la manipulation technique du contenu. Google a mis en place des politiques supplémentaires contre les fausses déclarations, y compris pour les annonceurs, qui examinent le comportement des créateurs de contenu. Ces règles pour les annonceurs s'appliquent également sur YouTube. Nous recherchons toujours de nouvelles menaces potentielles liées aux dommages personnels ou sociétaux résultant des nouvelles technologies, y compris celle-ci, et nous pourrions actualiser nos politiques à l'avenir si nous identifions des lacunes qui ne sont pas actuellement couvertes par nos règles ou systèmes existants. Par exemple, nous avons récemment mis à jour notre politique sur les images pornographiques involontaires (communément appelées «pornographie de vengeance») afin de couvrir les images factices en plus des images réelles."*

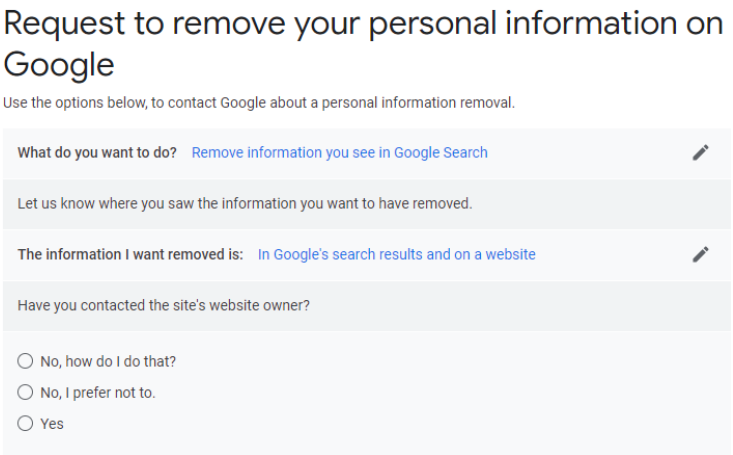
We take the threat of manipulated media very seriously, whether they are AI-generated or low-tech edits. YouTube has clear policies that outline what content is not acceptable to post and we remove videos violating these policies when flagged to us. We are always working to invest in and improve on our processes and technology to enforce our guidelines, including against potential threats related to synthetic media and disinformation.

YouTube's Community Guidelines prohibit certain deceptive practices that aim to take advantage of the YouTube community, including in some contexts those involving the technical manipulation of content. Google has additional policies in place against misrepresentation, including for advertisers, which looks at the behavior of content creators. These advertiser policies also apply on YouTube.

We are always looking into new potential threats related to personal or societal harm arising from new technologies, including this one, and may further update our policies in the future if we identify gaps that are not currently covered by our existing rules or systems. For example, we recently updated our Google policy on involuntary pornographic imagery (colloquially referred to as 'revenge porn') to cover fake imagery in addition to real imagery.

Source : capture extraite de la réponse de Google à Adam B. Schiff

Google offre également la possibilité aux utilisateurs de retirer des images ou vidéos *deepfakes* de ses résultats de recherche (mais pas du réseau internet) en remplissant un formulaire en ligne[61]. Il est cependant possible que le volume de ces documents soit trop important pour rendre cette méthode efficace, en particulier lorsqu'il s'agit de personnalités publiques ou célèbres.



Capture du formulaire Google pour demander le retrait d'un contenu

[61] Google, *Demande de suppression de vos informations personnelles de Google*, https://support.google.com/websearch/troubleshooter/9685456#ts=2889054%2C2889099

Google contribue aussi à lutter *technologiquement* contre les *deepfakes*. En effet l'entreprise distribue déjà un outil expérimental pour détecter les photos *deepfakes* (mais pas encore pour les vidéos)[62]. Et en 2019, Google a publié une base de données comprenant 3000 vidéos *deepfakes* (que Google enrichira au fur et à mesure) pour encourager et aider la recherche en matière de détection de ce type de contenu. En particulier, ces vidéos permettent d'entraîner des algorithmes de détection des *deepfakes*.

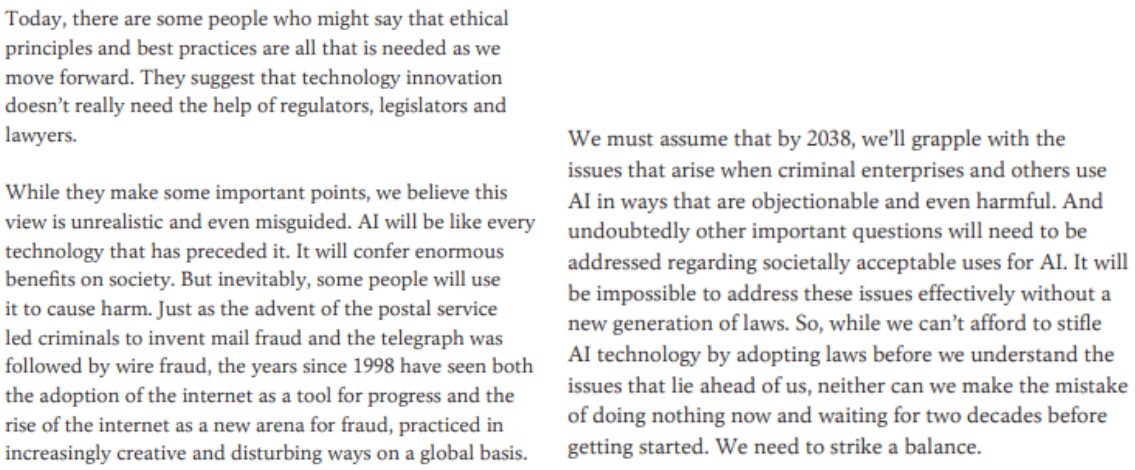
2.4.4. Microsoft

En matière de régulation des systèmes d'intelligence artificielle, et pour lutter contre les contenus *deepfakes*, Microsoft a adopté les principes éthiques suivants :

- **Justice** : les systèmes d'IA devraient traiter tout le monde équitablement.
- **Inclusivité** : les systèmes d'IA devraient responsabiliser tout le monde et impliquer les individus.
- **Fiabilité et sécurité** : les systèmes d'IA doivent fonctionner de manière fiable et sûre.
- **Transparence** : les systèmes d'IA doivent être compréhensibles.
- **Confidentialité et sécurité** : les systèmes d'IA doivent être sécurisés et respecter la vie privée.
- **Responsabilité** : les systèmes d'IA devraient avoir une responsabilité algorithmique.

Le président de Microsoft demande aux pouvoirs publics de réglementer l'intelligence artificielle, en particulier la reconnaissance faciale qui permet d'identifier les individus dans les lieux publics[63]. Dans un rapport de 2018, il écrivait ceci :

[62]https://www.technologyreview.com/2020/02/05/349126/google-ai-deepfakes-manipulated-images-jigsaw-assembler/
[63] Nitasha Tiku, *Microsoft Wants to Stop AI's 'Race to the Bottom'*, 2018, https://www.wired.com/story/microsoft-wants-stop-ai-facial-recognition-bottom/



Today, there are some people who might say that ethical principles and best practices are all that is needed as we move forward. They suggest that technology innovation doesn't really need the help of regulators, legislators and lawyers.

While they make some important points, we believe this view is unrealistic and even misguided. AI will be like every technology that has preceded it. It will confer enormous benefits on society. But inevitably, some people will use it to cause harm. Just as the advent of the postal service led criminals to invent mail fraud and the telegraph was followed by wire fraud, the years since 1998 have seen both the adoption of the internet as a tool for progress and the rise of the internet as a new arena for fraud, practiced in increasingly creative and disturbing ways on a global basis.

We must assume that by 2038, we'll grapple with the issues that arise when criminal enterprises and others use AI in ways that are objectionable and even harmful. And undoubtedly other important questions will need to be addressed regarding societally acceptable uses for AI. It will be impossible to address these issues effectively without a new generation of laws. So, while we can't afford to stifle AI technology by adopting laws before we understand the issues that lie ahead of us, neither can we make the mistake of doing nothing now and waiting for two decades before getting started. We need to strike a balance.

Source : Microsoft, *The Future Computed*, 2018 [64]

En vue des élections présidentielles américaines de Novembre 2020, Microsoft a déployé de nouveaux outils pour lutter contre les *deepfakes*[65]. Un outil nommé *Video Authenticator* permet aux médias et équipes de campagne de reconnaître les *deepfakes*. Un autre permet de tracer les documents grâce un à un système de marquage digital et de certificat d'authenticité (*Azure Information Protection*). Ces technologies étant nouvelles, il est difficile de savoir si elles auront un impact significatif. Par exemple, il n'est pas certain que nombre d'utilisateurs installent volontairement les extensions de navigation internet suggérées par Microsoft.

---

[64] Brad Smith & Harry Shum, *Artificial Intelligence and its role in society*, 2018, https://news.microsoft.com/uploads/2018/02/The-Future-Computed_2.8.18.pdf
"Aujourd'hui, il y a des gens qui pourraient dire que les principes éthiques et les bonnes pratiques sont tout ce dont nous avons besoin pour aller de l'avant. Ils suggèrent que l'innovation technologique n'a pas vraiment besoin de l'aide des régulateurs, des législateurs et des avocats. Bien qu'il soulève des points importants, nous pensons que ce point de vue est irréaliste et même erronée. L'IA sera comme toutes les technologies qui l'ont précédée. Elle apportera d'énormes avantages pour la société. Mais inévitablement, certaines personnes l'utiliseront à mauvais escient. Tout comme l'avènement du service postal a conduit les criminels à inventer la fraude postale et que le télégraphe a été suivi par la fraude du fil ('wire fraud'), on a vu depuis 1998 à la fois l'adoption d'Internet comme outil de progrès, mais également l'essor d'Internet en tant que nouvelle arène pour la fraude, pratiquée de façon de plus en plus créative et inquiétante à l'échelle mondiale. Nous devons supposer qu'en 2038, nous nous attaquerons aux problèmes qui surviennent lorsque des entreprises criminelles et d'autres utiliseront l'IA d'une manière qui est répréhensible et même nuisible. Et sans aucun doute, d'autres questions importantes devront être abordées concernant les utilisations socialement acceptables de l'IA. Il sera impossible de résoudre efficacement ces problèmes sans une nouvelle génération de lois. Ainsi, si nous ne pouvons pas nous permettre d'étouffer la technologie de l'IA en adoptant des lois avant de comprendre les problèmes qui nous attendent, nous ne pouvons pas non plus commettre l'erreur de ne rien faire maintenant et d'attendre deux décennies avant de commencer. Nous devons trouver un équilibre."
[65] https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/

En parallèle du lancement de ces nouveaux outils, Microsoft a annoncé rejoindre plusieurs initiatives :

- Un partenariat avec l'AI Foundation, pour mettre Video Authenticator à disposition des organisations via son programme Reality Defender 2020. L'enjeu est de mieux guider les organisations dans les limites et considérations éthiques relatives aux technologies de détection deepfake.
- Un autre partenariat avec un consortium de sociétés de médias, sous le nom de Project Origin, permettra de tester la technologie d'authenticité de Microsoft. Une initiative de plusieurs éditeurs et sociétés de médias sociaux, la Trusted News Initiative, a également accepté de s'engager avec Microsoft pour tester sa technologie.
- Un accord avec l'Université de Washington, Sensity et USA Today a également été conclu pour étoffer la littérature média et aider les individus à mieux identifier la désinformation et gérer les risques liés aussi bien au *deepfakes* qu'aux *cheap fakes*[66]. Dans le cadre de ce partenariat, une campagne de messages d'intérêt public encourage la population à prendre une « pause de réflexion ». En outre, la campagne invite les citoyens à vérifier que les informations proviennent d'organismes de presse réputés avant de les partager sur les médias sociaux. Et ce, en particulier en période électorale.

Par ces initiatives et ces partenariats, Microsoft met en avant l'éducation pour aidera les gens à trier la désinformation des faits réels.

2.4.5. Plateformes asiatiques

Cette attitude à l'égard des contenus *deepfakes* est cependant loin d'être universelle parmi les plateformes. Par exemple, Snapchat et TikTok[67] incorporent les technologies *deepfakes* dans leurs plateformes respectives, notamment pour proposer des filtres à leurs utilisateurs afin de modifier

[66] Le monde Informatique, *Microsoft dévoile des outils pour lutter contre les deepfakes*, 2020
[67] Michael Nuñez, *Snapchat and TikTok Embrace 'Deepfake' Video Technology Even As Facebook Shuns It*, 2020, https://www.forbes.com/sites/mnunez/2020/01/08/snapchat-and-tiktok-embrace-deepfake-video-technology-even-as-facebook-shuns-it/#7782d70042c0

leurs vidéos. C'est d'ailleurs en ce sens que Snapchat a acquis AI Factory[68], une entreprise spécialisée dans la création de *deepfakes*, basée en partie en Ukraine. Néanmoins, si ces plateformes adoptent les technologies *deepfakes*, elles mettent en place des dispositifs pour que leurs utilisateurs puissent facilement saisir qu'il ne s'agit pas de vraies vidéos. Il est d'ailleurs intéressant de noter que la version chinoise de TikTok, *Douyin*, implémente les normes chinoises plus strictes en la matière (par exemple labellisation des *deepfakes*, identités réelles des utilisateurs, protection de données). Ce n'est pourtant pas le cas de TikTok[69]. Cela tend à illustrer un manque de volontarisme chez certaines plateformes vis-à-vis de la protection contre un mauvais usage des *deepfakes*. A l'image de Bytedance (maison-mère de TikTok et Bayin), les plateformes peuvent faire le choix de s'adapter aux normes des différents pays et de m'imposer des barrières que si elles y sont contraintes.

❝

Cela tend à illustrer un manque de volontarisme chez certaines plateformes vis-à-vis de la protection contre un mauvais usage des deepfakes.

Pour d'autres, la technologie *deepfake* est à la base de leur service. C'est par exemple le cas de Zao, une application chinoise sortie en août 2019 qui s'était très rapidement retrouvée parmi les applications les plus téléchargées[70]. Cette application permet, grâce à des algorithmes de *deepfakes*, d'incruster son visage dans une séquence vidéo, notamment des films. Cette application avait cependant soulevé de nombreuses critiques à cause de ses conditions générales

[68] Ingrid Lunden, *Snapchat quietly acquired AI Factory, the company behind its new Cameos feature, for $166M*, 2020, https://techcrunch.com/2020/01/03/snapchat-quietly-acquired-ai-factory-the-company-behind-its-new-cameos-feature-for-166m/

[69] Josh Constine, ByteDance & TikTok have secretly built a deepfakes maker, 2020 https://techcrunch.com/2020/01/03/tiktok-deepfakes-face-swap/?guccounter=1

[70] Pierrine Sigoret, Ce qu'il faut savoir sur Zao, l'application au succès fulgurant qui vous transforme en Leonardo DiCapri, 2019, https://www.numerama.com/tech/545092-ce-quil-faut-savoir-sur-zao-lapplication-au-succes-fulgurant-qui-vous-transforme-en-leonardo-dicaprio.html

d'utilisation qui stipulaient que Zao conservait un droit permanent et irrévocable sur les photos fournies par les utilisateurs. Ces questions de droit à l'image soulèvent des questions dans une société asiatique qui utilise de plus en plus les caractéristiques du visage pour proposer divers services, comme le paiement par reconnaissance faciale. Ces technologies, encore peu démocratisées en Occident, pourraient rapidement s'y développer. C'est pourquoi il est nécessaire que les réglementations des pays occidentaux anticipent la potentielle émergence de ces nouveaux services.

La société Tencent qui est derrière l'application WeChat voit de nombreuses opportunités grâce au *deepfake,* qu'elle renomme *deep synthesis*[71]. Le laboratoire Tencent AI Lab en explore les applications dans la création de contenu, le marketing, les médias sociaux, le divertissement, le commerce électronique ou encore la communication. Par exemple, les chercheurs du Tencent AI Lab ont récemment publié un article montrant qu'il était possible de créer une voix chantante à partir d'échantillons d'une voix normale[72]. Il est donc clair que les entreprises chinoises vont continuer à investir dans ce domaine.

Naturellement, les pouvoirs publics sont plus circonspects. C'est ainsi qu'au 1er janvier 2020, l'autorité de régulation de l'internet en Chine, consciente que les *deepfakes* peuvent "*mettre en danger la sécurité nationale, perturber la stabilité sociale, perturber l'ordre social et porter atteinte aux droits et intérêts légitimes d'autrui*", a rendu obligatoire le marquage et signalement des documents *deepfakes* et de réalité virtuelle par leurs auteurs sous peine de sanctions.

Les plateformes internet ne sont cependant pas les seuls acteurs à prendre en compte dans ce débat. D'autres parties prenantes peuvent avoir un rôle à jouer.

[71] https://chinai.substack.com/p/chinai-104-tencent-2020-ai-white
[72] https://www.arxiv-vanity.com/papers/1912.10128/

2.5. Au-delà des acteurs de l'internet

De nombreuses questions éthiques vont trouver des réponses dans les principes adoptés par les acteurs de l'internet. Mais comme le soulignent le président de Microsoft et d'autres leaders du secteur, ces principes ne seront pas suffisants. De nouvelles lois et réglementations sont nécessaires, d'autant plus que certaines plateformes affichent un manque de volontarisme à s'auto-réguler.

Mais au-delà de ces nouvelles lois, d'autres acteurs sont amenés à être impliqués. Citoyens, employés, consommateurs et investisseurs peuvent aussi agir à leur échelle pour limiter les effets néfastes des infractions éthiques. Dans leur ensemble, ils peuvent notamment valoriser les entreprises les plus actives à résoudre les problèmes éthiques engendrés par leurs produits et services. Les employés peuvent partager leurs préoccupations avec le management, refuser de travailler sur des projets non-éthiques, ou décider de quitter l'entreprise. Les consommateurs peuvent boycotter des entreprises qui ont des pratiques éthiques douteuses. Les investisseurs peuvent demander que les considérations éthiques soient prises au sérieux par les entreprises dans lesquelles ils investissent, ou orienter leurs investissements vers des entreprises qui mettent davantage l'accent sur ces questions.

Toutes les parties prenantes ont leur rôle à jouer dans la lutte contre la mauvaise utilisation des algorithmes de génération de contenus.

Proposition n°12
Impliquer les employeurs dans la lutte contre les *deepfakes*. Les inciter à adopter des chartes éthiques, à informer leurs employés, clients, fournisseurs et partenaires de leurs bonnes pratiques éthiques sur le sujet. Encourager la mise en place de formations autour des *deepfakes*.

3. Cadre légal, régulation et gouvernance

3.1. Cartographie des acteurs

Les défis juridiques que posent les algorithmes de génération de contenu ne sont pas nouveaux et relèvent de champs bien établis du droit : liberté d'expression, droit à l'image ou droit de propriété notamment. Néanmoins, les outils juridiques ne semblent pas toujours adaptés face à l'évolution technologique (vitesse de propagation, volume, faible coût de production).

Quatre grands acteurs peuvent être cartographiés :

1. Les créateurs des algorithmes (entreprises, développeurs indépendants, etc.).
2. Les auteurs de contenus *deepfakes* (contenus basés sur les algorithmes).
3. Les plateformes internet (Facebook, Youtube, Reddit entre autres).
4. Les utilisateurs (qui regardent et partagent les *deepfakes* en ligne).

Cette liste permet d'évaluer les différents degrés de réglementation auxquels il faut légiférer pour encadrer les algorithmes de génération de contenu. Par ailleurs, si le droit encadrant les producteurs de contenu peut relever de l'échelle nationale, voire européenne, les enjeux liés aux opérations de déstabilisations menées par des Etats ou la réglementation des plateformes, dont la force de frappe est mondiale, nous semblent devoir relever de la sphère supra-étatique.

3.2. Cadre juridique

Il existe un cadre juridique préexistant et applicable aux algorithmes de génération de contenu. Il s'agit des règles traditionnelles relatives à la liberté d'expression, au droit à l'image et au droit de propriété auxquels se sont ajoutés plus récemment les lois protégeant les données personnelles et luttant contre les fausses nouvelles.

- **Liberté d'expression** : les algorithmes de génération de contenu constituent un nouvel outil d'expression permettant, au choix, de créer des œuvres d'art ou de diffuser des idées politiques ou humoristiques. Les messages qu'ils diffusent sont donc protégés par la liberté d'expression, liberté fondamentale consacrée au plan international (par la

Déclaration Universelle des Droits de l'Homme de 1948), au plan européen (par la Convention Européenne des Droits de l'Homme) et au plan national (par la Déclaration des Droits de l'Homme et du Citoyen de 1789). Cette liberté implique que tout individu a le droit d'avoir son opinion et ses idées et peut les exprimer par n'importe quel moyen, sous n'importe quel format et sur n'importe quel support : oral, écrit, audiovisuel, culturel, virtuel, artistique, etc. Bien qu'érigée au rang de "*droit universel qui constitue un élément de toute démocratie*"[73], la liberté d'expression peut néanmoins être limitée lorsque entrent en jeu :

- La sauvegarde de l'ordre public (sécurité nationale, intégrité territoriale, sûreté publique, prévention des infractions, protection de la santé publique, protection de l'ordre démocratique, protection de la morale, lutte contre le racisme, etc.).
- Les droits d'autrui (protection de la vie privée, protection contre la diffamation et l'injure, droit de réponse, informations confidentielles et secret des affaires, droit à l'image, présomption d'innocence, etc.).

- **Droit à l'image :** les algorithmes de génération de contenu sont souvent à l'origine de la diffusion d'images (photos ou vidéos). Entre alors en ligne de compte le droit à l'image, qui a été développé par la jurisprudence sur le fondement du droit au respect de la vie privée. Il s'agit du droit exclusif que toute personne a sur son image et l'utilisation qui en est faite, quel que soit son support (photos, vidéos, etc.) ou son contexte (vacances, événement familial, manifestation culturelle ou religieuse, etc.). Ce droit (tout comme la liberté d'expression) connaît également des limites, dont les principales sont :
  - Image non cadrée d'une personne prise dans un public.
  - Le droit à l'information, lorsque l'image illustre des événements d'actualité, un débat général sur la démocratie ou un phénomène de société, un sujet historique ou les fonctions d'une personnalité publique, sous réserve du respect de la dignité et de la vie privée des personnes.

[73] Conseil Constitutionnel 1994 : la liberté d'expression est une "*liberté fondamentale d'autant plus précieuse que son existence est une des garanties essentielles du respect des autres droits et libertés.*"

- La caricature ou le pastiche, sous réserve notamment du respect des lois du genre.

Mais en dehors de ces exceptions, l'auteur du contenu doit, avant toute diffusion, obtenir l'accord préalable de la personne représentée, en précisant l'usage précis pour lequel l'accord est donné. À noter que cette autorisation est d'interprétation stricte, le consentement d'une personne à être photographiée n'emportant pas pour autant son consentement à la diffusion de son image. Ces règles s'appliquent aux utilisateurs d'algorithmes de génération de contenu. Ainsi, l'auteur d'un algorithme de génération de contenu montrant l'image d'une personne, sans avoir recueilli l'assentiment de la personne représentée pour cet usage, engage sa responsabilité civile, sur le fondement de l'article 9 du Code civil. Il risque en outre d'être sanctionné pénalement (jusqu'à 1 an d'emprisonnement et 45 000 euros d'amende) si les infractions définies aux articles 226-1 (atteinte à l'intimité de la vie privée par la fixation de l'image d'une personne se trouvant dans un lieu privé) et 226-2 (communication au public de l'image d'une personne enregistrée sans son autorisation) du Code pénal sont reconnues.

- **Droit de propriété** : les algorithmes de génération de contenu posent la question de la propriété : qui est propriétaire du résultat créé et diffusé ?

  Dans l'affaire du tableau Bellamy (une peinture réalisée par une intelligence artificielle), s'est posée les questions de la titularité des droits d'auteur sur le tableau et de sa propriété. Qui était-donc l'auteur et le propriétaire de l'œuvre ?

  - Ian Goodfellow, inventeur des algorithmes de génération de contenu, dont l'équation mathématique figure au bas du tableau ?
  - Le créateur du code utilisé qui est accessible en open source sur internet ?
  - Les membres du collectif Obvious ?

  Nous avons eu l'opportunité de nous entretenir avec Hugo Caselles-Dupré, membre du collectif Obvious. Ce dernier nous a indiqué que si Ian Goodfellow n'avait jamais rien revendiqué, le créateur du code open source utilisé avait, quant à lui, tenté de revendiquer

la titularité du code et partant, du tableau. En l'espèce, un consensus sur la propriété des 11 tableaux de la série Bellamy a été trouvé entre le collectif et le créateur du code open source aux termes duquel la propriété et la titularité de l'œuvre ont été reconnues au collectif Obvious. Toutefois, les tribunaux n'ont pas eu à juger cette affaire, de sorte que la question de la titularité des droits d'auteur et de la propriété des œuvres créées via un système d'intelligence artificielle reste, à date, un débat essentiellement doctrinal. Les réflexions en cours (Résolution du 16 février 2017 du Parlement européen, consultation publique sur l'intelligence artificielle lancée par l'OMPI en décembre 2019) privilégient la protection de ces œuvres par le droit d'auteur, ce qui nécessiterait d'adapter ce droit, à ce jour principalement attaché à la personne de l'auteur. Il est à noter qu'un premier pas en ce sens a été fait en Chine, la juridiction chinoise de Nashan ayant récemment tranché un litige relatif à un système intelligent d'écriture et consacré la protection des œuvres générées via ce système sur le fondement du droit d'auteur (Nanshan District People's Court, Shenzhen Tencent Computer System Co., Ltd c. Shanghai Yingxun Technology Co., Ltd, 24 décembre 2019).

- **Protection des données à caractère personnel :** Depuis l'entrée en application, le 25 mai 2018, du Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données (ci-après "Règlement Général sur la Protection des Données" ou "RGPD"), les traitements[74] de données à caractère personnel sont licites à condition d'être mis en œuvre sur la base de l'un des fondements légaux listés à l'article 6 de ce règlement (notamment l'intérêt légitime du responsable de traitement, l'exécution d'un contrat ou consentement des personnes concernées).

  Les responsables de traitement doivent en outre respecter les principes suivants :

[74] "toute opération ou tout ensemble d'opérations effectuées ou non à l'aide de procédés automatisés et appliquées à des données ou des ensembles de données à caractère personnel, telles que la collecte, l'enregistrement, l'organisation, la structuration, la conservation, l'adaptation ou la modification, l'extraction, la consultation, l'utilisation, la communication par transmission, la diffusion ou toute autre forme de mise à disposition, le rapprochement ou l'interconnexion, la limitation, l'effacement ou la destruction" (article 4 RGPD)

- Principes de licéité, de loyauté et de transparence (les traitements doivent s'inscrire dans l'un des fondements légaux listés par le RGPD et les informations relatives aux traitements doivent être facilement accessibles et formulées en termes clairs et simple).
- Principe de limitation des données, lesquelles doivent être adéquates, pertinentes et limitées à ce qui est nécessaire au traitement.
- Principe d'exactitude des données.
- Principe de limitation dans le temps de la conservation des données.
- Principe de sécurité des données, notamment contre tout risque de traitement non licite, perte, destruction ou dégât accidentel ou volontaire.

Le RGPD prévoit en outre, sauf limitations éventuelles, les droits suivants pour les personnes concernées :

- Le droit d'être informé sur les caractéristiques du traitement.
- Le droit d'accès aux données et de leur rectification si elles sont inexactes.
- Le droit à l'effacement des données (ou "droit à l'oubli"), si le traitement était fondé sur le consentement de la personne et que celle-ci le retire ou lorsque les données ne sont plus nécessaires au regard des finalités du traitement).
- Le droit à la limitation du traitement, lorsque la personne concernée conteste l'exactitude des données, pendant le temps nécessaire à la vérification de cette exactitude.
- Le droit à la portabilité des données, afin de permettre à la personne concernée de recevoir ses données dans un format structuré, couramment utilisé et lisible par machine, afin de les transmettre à un autre responsable Le traitement (par exemple pour changer de prestataire de services).
- Le droit d'opposition, notamment si le traitement vise les actions de profilage ou de prospection commerciale.

- Le droit de ne pas faire l'objet d'une décision individuelle automatisée (sauf si cette décision est nécessaire à la conclusion ou l'exécution d'un contrat ou si la personne concernée y a consenti explicitement).

- **Protection contre les fausses nouvelles :** l'article 27 de la loi du 29 juillet 1881 sur la liberté de la presse prévoit que la publication, la diffusion ou la reproduction de nouvelles fausses et/ou de pièces fabriquées, falsifiées ou mensongèrement attribuées à des tiers est punie d'une amende de 45 000 euros, lorsque cette action est faite de mauvaise foi et qu'elle trouble ou est susceptible de troubler la paix publique.

  Il s'agit toutefois d'un "délit contre la chose publique" et non d'un délit contre les personnes, de sorte que seules les entités ayant un intérêt à la défendre peuvent s'en prévaloir (le Ministère Public pour le compte de l'Etat mais également des associations ayant pour objet la défense de la chose publique par exemple). Un particulier ne pourrait en revanche a priori pas déposer de plainte ou engager une action judiciaire sur le fondement de cette infraction.

  Par ailleurs, depuis le 22 décembre 2018, la loi n° 2018-1202 relative à la lutte contre la manipulation de l'information (appelée couramment "loi anti *fake news*") prévoit les mesures suivantes :

  - L'obligation pour les plateformes de dévoiler l'identité des auteurs des publicités diffusées en période électorale, ainsi que les montants consacrés à ces publicités. Cette obligation de transparence n'est toutefois soumise à aucune sanction.
  - La possibilité de saisir en référé (procédure d'urgence) le tribunal judiciaire, pendant les trois mois précédant le premier jour du mois d'élections générales et jusqu'à la date du tour de scrutin où celles-ci sont acquises, pour lui demander de prendre les mesures nécessaires pour faire cesser la diffusion des informations. Le juge statue sous 48 heures si (i) les fausses informations diffusées le sont "de manière artificielle ou automatisée" et "massive", si (ii) le caractère inexact ou

trompeur de l'information est manifeste et si (iii) le risque d'altération de la sincérité du scrutin est également manifeste.

- Le pouvoir pour le Conseil Supérieur de l'Audiovisuel (CSA) d'empêcher ou de suspendre la diffusion des services de télévision contrôlés par un Etat étranger ou sous l'influence de cet Etat, et portant atteinte aux intérêts fondamentaux de la nation. Le CSA peut également émettre des recommandations, effectuer des contrôles, et éventuellement sanctionner les plateformes ne se conformant pas à l'obligation de coopération définie ci-dessous.
- L'obligation pour les opérateurs de plateforme en ligne de mettre à la disposition de leurs utilisateurs un dispositif facilement accessible et visible permettant de signaler de fausse information, ainsi que de mettre en œuvre des mesures complémentaires, pouvant notamment porter sur la transparence de leurs algorithmes, la lutte contre les comptes propageant des fausses informations ou l'information des utilisateurs sur la nature, l'origine et les modalités de diffusion des contenus, ainsi que sur l'identité des personnes qui les rémunèrent en contrepartie de la promotion de contenus d'information se rattachant à un débat d'intérêt général. Ces mesures doivent être rendues publiques.

3.3. Quels outils le droit français offre-t-il aujourd'hui pour faire face à des algorithmes de génération de contenu attentatoires aux droits des personnes ?

La France ne s'est pas dotée de loi spécifique sur les algorithmes de génération de contenu, mais les règles existantes permettent de réglementer certaines dérives engendrées par ces algorithmes et leurs possibles atteintes aux droits des personnes. On pense plus particulièrement aux *deepfakes*, évoqués précédemment. Dans ce cadre, l'image d'une personne identifiable est utilisée et, suivant le cas, peut être attentatoire à son droit à l'image, diffamatoire, injurieuse et/ou véhiculer de fausses informations.

Dans ce cadre, la victime dispose de deux options :

- **Agir contre la plateforme,** en référé, suivant la voie qui lui est offerte depuis la loi du 21 juin 2004 sur la confiance pour l'économie numérique, afin de solliciter toutes mesures propres à prévenir un dommage ou à faire cesser un dommage occasionné par le contenu d'un service de communication au public en ligne. Le plus fréquemment la demande portera sur le retrait de la publication litigieuse.

  L'article 6 I de la loi n° 2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique (dite "loi LCEN") prévoit en effet la possibilité pour toute personne :

  - De notifier à l'hébergeur d'un site internet une demande de retrait d'un contenu litigieux dudit site, sous réserve de décrire ledit contenu et sa localisation précise et d'avoir préalablement tenté de joindre son auteur (article 6 I 5);
  - D'engager une action en référé ou de déposer une requête pour demander à un prestataire d'hébergement ou fournisseurs d'accès à internet de prendre "toutes mesures propres à prévenir un dommage ou à faire cesser un dommage occasionné par le contenu d'un service de communication au public en ligne", notamment le retrait dudit contenu ("référé-internet" prévu à l'article 6 I 8).

- **Agir contre l'éditeur du contenu** et solliciter des dommages et intérêts aux fins de réparer son préjudice. Ce sont ces différents cas que nous allons parcourir et illustrer dans la partie qui suit.

3.3.1. Cas n°1 : *deepfake* attentatoire au droit à l'image

Il s'agit du cas où l'image d'une personne est utilisée sans son autorisation.

Dans un tel cas, la personne dont l'image a été utilisée peut :

- Contacter le responsable du site sur lequel est publiée l'image et lui en demander la dépublication, sur le fondement de son droit à l'effacement des données à caractère personnel la concernant. À défaut d'une réponse satisfaisante dans un délai de 2 mois cette personne peut saisir gratuitement la CNIL qui pourra infliger plusieurs types de sanctions (avertissement, injonctions, sanctions pécuniaires...).
- Saisir les juridictions civiles, le cas échéant en référé (c'est-à-dire en urgence) afin d'obtenir le retrait des images litigieuses et l'octroi de dommages et intérêts en réparation des préjudices subis (article 9 du Code civil). La décision n'est pour autant souvent obtenue qu'après plusieurs semaines.
- Porter plainte contre l'auteur de cette diffusion, sur le fondement des articles 226-1 à 226-2 du Code pénal.

3.3.2. Cas n°2 : *deepfake* diffamatoire (articles 29 et suivants de la loi du 29 juillet 1881 sur la liberté de la presse)

Le *deepfake* est diffamatoire lorsqu'il allègue ou impute un fait relatif à la vie privée d'une personne, qui porte atteinte à son honneur ou à sa considération.

Dans le cadre des exemples de *deepfakes* cités précédemment, pourraient être considérées comme diffamatoires les vidéos représentant Mark Zuckerberg et Nancy Pelosi.

La victime pourra alors porter plainte contre l'auteur de la diffusion (qui encourt, lorsque la diffusion est publique, une amende pouvant aller jusqu'à 12 000 € ou 45 000 € suivant les cas, pour une diffamation commise envers des particuliers) et solliciter des dommages et intérêts pour réparer le préjudice subi du fait de la publication diffamatoire.

3.3.3. Cas n°3 : *deepfake* injurieux (articles 33 et suivants de la loi du 29 juillet 1881 sur la liberté de la presse)

Le *deepfake* est injurieux lorsque l'accusation qu'il contient n'est pas un fait vérifiable et que son but est de blesser ou d'offenser la personne représentée. Dans ce cadre, l'auteur du *deepfake* ne pourra pas invoquer sa bonne foi pour s'exonérer.

Comme pour le fait diffamatoire, la victime devra porter plainte et l'auteur encourt les mêmes sanctions.

3.3.4. Cas n°4 : *deepfake* diffuseur de fausses nouvelles (article 27 de la loi du 29 juillet 1881 sur la liberté de la presse et loi "anti *fake news*" du 22 décembre 2018)

Classiquement, on définit la fausse nouvelle comme celle (potentiellement diffamatoire, injurieuse, troublant l'ordre public, dénigrante etc.) dont l'objectif peut être, suivant le cas, de manipuler l'opinion (ex. lobbies, groupes de pression), de fausser un cours de bourse, de discréditer un concurrent, d'orienter des choix politiques, de distiller de la haine contre une communauté ou encore d'attirer du trafic en diffusant des informations extravagantes ou provocatrices.

- Lorsque le *deepfake* est (i) le vecteur de diffusion d'une fausse nouvelle, de pièces fabriquées, falsifiées ou mensongèrement attribuées à des tiers et (ii) qu'il trouble ou est susceptible de troubler l'ordre public, il tombe sous le coup du délit de fausses nouvelles sanctionné par l'article 27 de la loi de 1881 sur la presse. Toutefois, comme cela a été expliqué ci-avant[75], seul le Ministère Public peut engager une action sur le fondement de cet article.
- Par ailleurs, le référé civil créé par la loi du 22 décembre 2018 relative à la lutte contre la manipulation de l'information (dite loi "anti-fake news"), permet de demander de faire cesser la diffusion de fausses informations durant les trois mois précédant un scrutin national. Quand il est saisi, le juge des référés doit apprécier, sous 48 heures, si (i) les fausses informations diffusées le sont "de manière artificielle ou automatisée" et "massive",

[75] titre 3.2 du présent rapport

si (ii) le caractère inexact ou trompeur de l'information est manifeste et si (iii) le risque d'altération de la sincérité du scrutin est également manifeste. Il est ici important de noter que malgré son libellé extrêmement large, la loi anti-fake news est strictement limitée aux périodes et conséquences électorales.

Mais si le *deepfake* est susceptible d'avoir des conséquences ou implications dans un secteur particulier, d'autres lois spécifiques sont susceptibles de s'appliquer :

- Le délit de pratique commerciale trompeuse (article L.121-2 à L.121-4 et L.132-2 du Code de la Consommation.
- La concurrence déloyale, en ce compris le parasitisme et le dénigrement.
- Le délit de "faits faux de nature à porter atteinte au crédit de la Nation" (article 410-1 du Code pénal).
- Le délit d'atteinte à la transparence du marché (articles L.465-3-1 et s. du Code monétaire et financier).
- Le montage réalisé avec les paroles ou l'image d'une personne sans son consentement (article 226-8 du code pénal).
- Le délit d'usurpation d'identité (article 226-4-1 du Code pénal).
- Le délit de dénonciation calomnieuse (article 226-10 du Code pénal).
- Le délit de divulgation de menaces et fausses alertes (article 322-14 du Code pénal).
- Le droit instauré par le RGPD de ne pas faire l'objet d'une décision exclusivement fondée sur un traitement automatisé de ses données, y compris le profilage, produisant des effets juridiques la concernant ou l'affectant de manière significative, sous réserve des exceptions visées ci-avant[76].

[76] Titre 3.2 du présent rapport

Le droit français n'est donc pas totalement démuni face aux *deepfakes* et possède plusieurs instruments permettant aux victimes de solliciter le retrait des publications frauduleuses et/ou - du moins en théorie face à la difficile identification des auteurs - la réparation du préjudice subi. Toutefois, peu de victimes connaissent l'étendue de leurs droits. Il conviendrait dès lors de développer l'existence et de renforcer la visibilité des structures d'accompagnement existantes.

Proposition n°7
Favoriser l'accès au droit des victimes de *deepfake*, d'usurpation d'identité ou de harcèlement (aux voies de recours et à l'information de leurs droits) par le développement et la visibilité des structures d'accompagnement (d'accueil physique ou plateformes étatiques).

On constate également que la plupart des règles précitées n'ont vocation à s'appliquer qu'au bénéfice des consommateurs d'algorithmes de génération de contenu (afin de les protéger) et à l'encontre des producteurs privés de ces algorithmes de génération de contenu (afin de les sanctionner). Peu de règles existantes (à l'exception de la loi pour la confiance dans l'économie numérique) visent directement les plateformes qui, comme leur nom l'indique, jouent un rôle central dans la diffusion à grande échelle des contenus (dont les algorithmes de génération de contenu).

3.4. Quelle doit être la responsabilité encourue par les plateformes ?

Les plateformes numériques jouent aujourd'hui un rôle majeur dans la diffusion de contenus. Les exemples présentés en première partie montrent à quel point les plateformes accélèrent la diffusion et multiplient le volume des contenus. Le débat sur la responsabilité des plateformes est extrêmement vif du fait de l'exemption légale dont elles font l'objet. La loi LCEN visée ci-dessus, qui transpose la directive 2000/31/CE dite "*e-commerce*", prévoit en effet un régime dérogatoire de responsabilité pour les prestataires d'hébergement, c'est-à-dire les prestataires qui assurent le stockage de contenus fournis par des tiers en vue de leur mise à disposition par des services de communication au public en ligne, ce qui inclut les plateformes sur lesquelles des contenus produits par des tiers sont diffusés.

Ces hébergeurs ne peuvent voir leur responsabilité pénale engagée du fait de contenus hébergés à la demande des producteurs de ces contenus, si elles n'avaient pas effectivement eu connaissance de l'activité ou de l'information illicites, notamment au moyen du système de notification prévu à l'article 6 I 5 de la loi LCEN, telle que décrite ci-avant, ou si, dès le moment où elles en ont eu connaissance, elles ont agi promptement pour retirer ces informations ou en rendre l'accès impossible. Ce régime dérogatoire s'applique aux plateformes, dès lors qu'elles ne sont pas éditrices de ces contenus, ne les ont pas déterminés ou sélectionnés et n'en ont pas eu connaissance ou ne les ont pas vérifiés *a priori*. Le législateur européen souhaitait développer la confiance en l'économie digitale par une meilleure information des utilisateurs et alléger les obligations des hébergeurs avec une réglementation plus avantageuse qui limite les coûts de mise en conformité.

Pour anticiper les réponses des pouvoirs publics, les plateformes de taille importante ont pris de manière volontaire des mesures préventives dont les initiatives sont détaillées dans l'encadré qui suit.

Mesures volontaires prises par les plateformes face au *deepfakes*

Facebook, Amazon et Microsoft ont lancé un concours de détection des *deepfakes*, le *Deepfake Detection Challenge* (DFDC) pour développer des algorithmes qui peuvent identifier des *deepfakes* et des médias qui ont été manipulés[77]. Facebook a également décidé de supprimer les documents *deepfake* de sa plateforme[78] sur la base des critères suivants :

- Le document a été édité ou synthétisé, au-delà des ajustements pour la clarté ou la qualité, d'une manière qui n'est pas évidente pour un individu moyen et qui pourrait induire en erreur en laissant penser qu'un sujet de la vidéo a dit des mots qu'il n'a pas réellement dit ;
- Le document est le produit de l'intelligence artificielle ou de l'apprentissage automatique qui fusionne, remplace ou superpose le contenu sur une vidéo, donnant ainsi l'impression qu'elle est authentique ;
- Les contenus parodiques ou satiriques sont exclus de restriction.

Twitter est également en train de mettre en place sa politique contre les *deepfakes*[79]. En particulier, Twitter signalera les utilisateurs des comptes qui partageront des documents *deepfake*, avertira les comptes qui souhaitent partager ces documents ("*retweet*") et fournira les raisons pour lesquelles il pense que le document est un *deepfake*.

Google offre la possibilité de retirer des images ou vidéos *deepfake* de ses résultats de recherche (mais pas du réseau internet) en remplissant un formulaire

[77] Facebook AI, *Deepfake Detection Challenge Dataset*, 2020, https://deepfakedetectionchallenge.ai/
[78] Facebook, *Enforcing Against Manipulated Media*, 2020, https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/
[79] Sarah Perez, *Twitter drafts a deepfake policy that would label and warn, but not always remove, manipulated media*, 2019, https://techcrunch.com/2019/11/11/twitter-drafts-a-deepfake-policy-that-would-label-and-warn-but-not-remove-manipulated-media/

en ligne[80]. Il est possible que le volume de demandes soit trop important pour rendre cette méthode efficace, notamment s'agissant de personnalités publiques ou célèbres.

Ces approches ne sont pas forcément suivies par toutes les plateformes. **Snapchat** et **TikTok**[81] comptent incorporer une fonctionnalité de création de *deepfakes* pour augmenter l'utilisation de leurs plateformes. Snapchat a notamment acquis AI Factory[82], une entreprise spécialisée dans la création de *deepfakes*, basée en partie en Ukraine.

Néanmoins, la réglementation des plateformes n'a cessé de progresser, notamment avec la mise en place du Règlement Général sur la Protection Données (RGPD) depuis le 25 mai 2018. Si le RGPD n'a pas vocation à encadrer les *deepfakes* ou les fausses nouvelles, il est néanmoins un outil efficace que les régulateurs peuvent utiliser pour juguler les contenus de *deepfakes*, en particulier ceux utilisant l'image de personnes dont le consentement n'a pas été recueilli. L'application Zao, devenue très populaire en Chine pendant l'été 2019, proposait des services de *deepfakes* en remplaçant les visages de stars par celui de l'utilisateur[83], mais ses conditions d'utilisation stipulaient que les droits consentis par les internautes sur leurs images étaient "*gratuits, irrévocables, et permanents*". Si ce genre de pratiques est courante avec les applications mobiles, il n'en est pas moins peu regardant des obligations du RGPD, notamment sur la finalité du traitement qui doit être clairement établie dans les conditions générales d'utilisation.

Le RGPD pourrait être néanmoins utilisé aux fins suivantes :

[80] Google Search Help, https://support.google.com/websearch/answer/9116649?hl=en
[81] Michael Nuñez, Snapchat and TikTok Embrace 'Deepfake' Video Technology Even As Facebook Shuns It, 2020, https://www.forbes.com/sites/mnunez/2020/01/08/snapchat-and-tiktok-embrace-deepfake-video-technology-even-as-facebook-shuns-it/#5995495242c0
[82] Ingrid Lunden, *Snapchat quietly acquired AI Factory, the company behind its new Cameos feature, for $166M*, 2020
[83] Alexis Zema, *Zao, l'application de vidéos «deepfake» qui inquiète les internautes chinois,* 2019, https://www.lefigaro.fr/secteur/high-tech/vie-privee-les-videos-deepfake-de-l-application-zao-inquietent-les-internautes-chinois-20190902

- Pour sanctionner les producteurs d'algorithmes de génération de contenu qui collectent illicitement les images de personnes
- Pour responsabiliser les plateformes qui diffusent les algorithmes de génération de contenu. Le Règlement général sur la protection des données (RGPD) repose sur le principe de la responsabilisation des acteurs, à défaut d'une déclaration ou d'une autorisation *a priori*, qui doivent mettre en place des programmes d'utilisation, de partage et de protection des données
- Pour utiliser les ressources actuelles de la CNIL à des fins d'éducation, d'audit et de mise en demeure. Le RGDP liste des obligations auxquelles les autorités nationales comme la CNIL doivent répondre. Entre autres, des obligations d'éducation, d'audit et de mise en demeure pour non-respect du règlement. Par exemple, la CNIL a produit des rapports sur les objets connectés et l'intelligence artificielle. Les algorithmes de génération de contenu automatisé pourraient aussi être étudiés et faire l'objet d'une discussion publique.

Si le RGPD peut être un outil de politique publique intéressant, il reste néanmoins limité pour plusieurs raisons :

- Il s'applique seulement aux traitements de données à caractère personnel.
- L'application du RGPD en dehors du territoire de l'Union Européenne est problématique. Si les autorités de régulation, comme la CNIL en France, disposent de certains moyens, les limites du principe d'extraterritorialité du RGPD sont aussi réelles. Dans le cas d'un acteur étranger utilisant des données de ressortissants européens en dehors de l'Union Européenne, la question peut se poser de savoir dans quelle mesure une décision d'une cour européenne est appliquée.

Un autre outil de politique publique plus efficace pourrait consister à responsabiliser davantage les plateformes.

La directive *e-commerce* de 2000, transposée dans le droit français en 2004 par la loi pour la confiance dans l'économie numérique, prévoit un régime juridique accommodant pour les

plateformes comme Facebook ou Google. Aujourd'hui, les obligations suivantes des plateformes ne semblent plus suffisantes face aux enjeux liés aux algorithmes de génération de contenu :

- Obligation de retirer "*promptement*" toute donnée dont le contenu serait "*manifestement illicite*".
- Obligation d'informer "*promptement*" les autorités publiques compétentes des signalements reçus par les utilisateurs pour des activités illicites limitées (incitation à la haine, proxénétisme, pédopornographie, détournement de mineurs). La plateforme gouvernementale Pharos (Plateforme d'harmonisation, d'analyse, de recoupement et d'orientation des signalements) comprend une équipe de 20 enquêteurs pour traiter les plaintes des plateformes mais aussi directement des usagers.
- Obligation d'un dispositif "*facilement accessible et visible permettant à toute personne de porter à leur connaissance*" les contenus contraires à l'intérêt général.
- Obligations de communiquer sur les moyens mis en œuvre dans la lutte contre les activités illégales.
- Obligations de transparence lorsqu'elles diffusent des contenus contre rémunération. Celles qui dépassent un certain volume de connexions par jour doivent avoir un représentant légal en France, mettre en œuvre des mesures pouvant porter sur la transparence de leurs algorithmes et publier des statistiques agrégées sur le fonctionnement de leurs algorithmes de recommandation, classement ou référencement de contenus d'information se rattachant à un débat d'intérêt général.

D'autres avis non contraignants ont été parallèlement émis par des autorités publiques. Ainsi, la recommandation du CSA (15 mai 2019) propose aux plateformes dépassant le seuil de 5 millions de visiteurs par mois des mesures concrètes destinées à assurer la transparence de l'information. De même, le rapport annuel de l'Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI), en avril 2019, met en garde contre les opérations de déstabilisation et d'influence.

Au vu de la place prise par les plateformes et leur importance dans la diffusion et la production de contenu (notamment par les algorithmes de sélection de contenu ou le design des plateformes), un statut intermédiaire pourrait permettre une meilleure prise en compte des

réalités actuelles. Ainsi la Commission Européenne est actuellement en train de discuter de mesures en ce sens.[84] Le Sénat français a quant à lui proposé un statut intermédiaire, celui d'éditeur de services, dans une proposition de résolution européenne en 2018. L'éditeur de services serait "*une société qui retire un avantage économique direct de la consultation des contenus hébergés*"[85], soumise à des obligations intermédiaires entre celles supportées par les éditeurs et celles des hébergeurs. Le tableau ci-dessous résume ce statut intermédiaire :

Editeur *(sans changement)*	Editeur de services *(nouvelle catégorie)*	Hébergeur *(sans changement)*
Régime de responsabilité de **droit commun** : - a une obligation d'identification des personnes qui ont créé un contenu qu'il héberge ; - a une obligation de surveillance du contenu des pages éditées : il s'agit d'une **obligation de résultat,** et non de moyens.	Régime de **responsabilité intermédiaire** L'éditeur de services : - a une obligation d'identification des personnes qui ont créé un contenu qu'il héberge ; - a une obligation de mettre en place les moyens, conformes à l'état de l'art, de surveillance des informations qu'il transmet ou stocke, et de recherche des faits ou des circonstances révélant des activités illicites ; il s'agirait donc d'une obligation de moyens, non de résultat ; - peut être tenu civilement ou pénalement responsable s'il a connaissance d'activités ou d'information manifestement illicites et qu'il n'agit pas promptement pour retirer ces informations ou en rendre l'accès impossible.	Régime de **responsabilité aménagée** L'hébergeur : - a une obligation d'identification des personnes qui ont créé un contenu qu'il héberge ; - n'a pas d'obligation générale de surveillance des informations qu'il transmet ou stocke, ni d'obligation générale de rechercher des faits ou des circonstances révélant des activités illicites ; - n'a pas d'obligation générale de surveillance des informations qu'il transmet ou stocke, ni d'obligation générale de rechercher des faits ou des circonstances révélant des activités illicites ; - ne peut être tenu civilement ou pénalement responsable que s'il a connaissance d'activités ou d'information manifestement illicites et qu'il n'agit pas promptement pour retirer ces informations ou en rendre l'accès impossible.

Source : Senat, *Lutte contre la contrefaçon : premier bilan de la loi du 29 octobre 2007*, 2020

On notera également la proposition de loi contre la haine sur internet, déposée le 20 mars 2019, qui proposait de mettre en place un nouveau régime de responsabilité administrative applicable aux opérateurs de plateformes à fort trafic. Cette proposition de loi suggérait d'imposer à ces opérateurs de plateforme, de retirer ou de rendre inaccessible, dans un délai maximal de 24 heures

[84] Julien Lausson, *Bruxelles entend accentuer la responsabilité judiciaire des géants du net*, 2020, https://www.numerama.com/politique/598046-bruxelles-entend-accentuer-la-responsabilite-judiciaire-des-geants-du-net.html

[85] Senat, *Lutte contre la contrefaçon : premier bilan de la loi du 29 octobre 2007*, 2020, http://www.senat.fr/rap/r10-296/r10-2969.html

après notification, tout contenu comportant manifestement une incitation à la haine ou une injure discriminatoire à raison de la race, de la religion, du sexe, de l'orientation sexuelle ou du handicap. Le manquement à cette obligation devait être passible d'une sanction susceptible d'atteindre 4 % du chiffre d'affaires annuel mondial de l'opérateur en cause. Cette proposition de loi a été censurée par le Conseil constitutionnel qui y a décelé de potentielles atteintes à la liberté d'expression.

Si les mesures éthiques et non-contraignantes prises par les grandes plateformes se rapprochent des critères de responsabilité intermédiaire listés ci-dessus, cela pourrait indiquer qu'un tel régime bénéficierait aux grands acteurs américains. Un système plus souple, celui de la responsabilité limitée, pourrait être envisagé pour les plateformes en dessous d'un certain seuil (en nombre d'utilisateurs, de trafic, etc.).

En plus des efforts d'autorégulation significatifs faits par les plateformes, des outils de politique publique existent, tel le RGPD, ou sont en voie de développement, comme la responsabilité intermédiaire. Les projets ci-dessous permettraient de renforcer davantage l'arsenal législatif actuel.

3.5. Les projets en cours

Le droit français se heurte en revanche principalement à deux problèmes majeurs :

- **L'inadéquation des sanctions pénales encourues par les auteurs au regard des conséquences subies sur l'ordre public** (sans même parler des conséquences sur les victimes mais que la sanction pénale n'a pas vocation à réparer).

  En effet, les sanctions prévues aujourd'hui sont principalement contenues dans des instruments législatifs datant du XIXème siècle, à l'heure où il était moins aisé d'avoir accès aux instruments de diffusion publique (journaux principalement, ce qui impliquait d'avoir un contact chez un éditeur) et surtout où la diffusion était bien moindre d'un point de vue

géographique (les journaux ayant des diffusions locales voire nationales pour quelques-uns).

À l'heure actuelle, où il suffit d'un smartphone et d'une connexion internet pour diffuser une image ou un contenu au monde entier en quelques secondes, l'auteur des diffusions illégales a moins la mesure de l'acte illégal qu'il commet. Les conséquences sur la tranquillité publique et sur la victime sont, elles, inversement proportionnelles aux démarches qu'il a fallu à l'auteur pour diffuser son contenu.

C'est pourquoi nous suggérons de renforcer les sanctions pénales encourues par les auteurs dans deux optiques : (i) un objectif de dissuasion et (ii) un objectif d'apaisement du trouble causé par les diffusions litigieuses à l'ordre public.

Proposition n°8
Renforcer les sanctions pénales existantes contre les auteurs de contenus de *deepfake* et les réévaluer en prenant en compte le nouvel environnement technologique et digital où la création de *deepfakes* est rendue possible**.**

Nous suggérons également, ainsi qu'il est proposé infra, en proposition 11, d'imposer aux éditeurs de contenus de signaler tout contenu deepfake afin de limiter l'impact pour les victimes, et qu'à défaut de signalement, les sanctions pénales encourues soient aggravées.

- **La responsabilité des diffuseurs.** L'identification des auteurs de ces nouvelles est complexe. C'est la raison pour laquelle les nouveaux instruments légaux émis visent à toucher les diffuseurs et non directement les auteurs. Cette voie nous paraît également la plus efficace pour lutter, non pas contre l'édition de ces contenus, ce qui est

nécessairement difficile à endiguer, mais contre leur diffusion au plus grand nombre, génératrice de préjudice.

Cette responsabilité des diffuseurs reste néanmoins, elle aussi, très théorique, eu égard :

- Au fait que la responsabilité des hébergeurs n'est engagée que s'ils n'ont pas "*promptement*" retiré une publication illicite ou bloqué l'accès d'un site qu'on leur avait signalé. Aucune sanction n'est donc encourue en cas de prompte obtempération mais surtout aucune définition quantitative n'est donnée de cette exigence de réactivité, pouvant donner lieu à une certaine souplesse de l'interprétation qui en est faite par les plateformes, susceptibles de se retrancher derrière des obstacles techniques.
- Aux moyens dont les plateformes disposent pour exercer un potentiel recours contre l'éditeur du contenu (celui qui est à l'origine de la publication et l'information – ex. auteur d'un tweet).

Nombreux sont donc ceux qui militent pour le renforcement de la responsabilité des grandes plateformes au vu notamment de leur puissance. À ce titre, il est intéressant de se pencher sur les systèmes mis en œuvre à travers le monde.

3.6. Exemples internationaux

Les *fake news* se diffusant partout dans le monde, la plupart des Etats semblent s'être emparés de ce problème, chacun trouvant son rempart "*anti-fake news*"[86].

[86]Corentin Lacoste, *À l'étranger, quelles législations contre les fake news ?*, 2018, https://www.liberation.fr/politiques/2018/06/09/a-l-etranger-quelles-legislations-contre-les-fake-news_1657644

3.6.1. En Europe

Plusieurs pays européens ont adopté des mesures dans le but de lutter contre la désinformation et plus précisément les *fake news*. Ainsi, à titre d'exemple :

- L'Allemagne a adopté une loi dénommée "la mesure NetzDG" prévoyant une amende de 50 millions € à l'encontre des réseaux sociaux qui ne suppriment pas les publications haineuses ou les f*ake news*.
- L'Angleterre s'est quant à elle distinguée en créant une agence spéciale, appelée "*National Security Communications Unit*" destinée à lutter contre les *fake news*.
- L'Italie a lancé en janvier 2018 un site internet, en lien avec la Police, sur lequel toute personne peut signaler les potentielles *fake news* qu'il aurait pu identifier.

À ce stade, seule la Suisse semble faire figure d'exception, concluant dans un rapport de mai 2017 que, "*faute de recul, il n'est cependant pas possible à l'heure actuelle d'affirmer dans quelle mesure une réglementation étatique est nécessaire*."

3.6.2. En dehors de l'Europe

De leur côté, les Etats-Unis ne disposent pas de loi fédérale sur les *deepfakes*, mais les discussions ont été soutenues à Washington au cours des deux dernières années[87]. Cependant certains Etats ont pris des mesures. On peut notamment citer l'État de Virginie, qui a adopté une loi en juillet 2019, incluant les "*deepfake*" dans sa lutte contre les "*revenge porn*" (pratique qui consiste à partager un contenu pornographique sans le consentement de la personne qui y apparaît). Ainsi les auteurs de ces partages (véridiques ou falsifiés via des algorithmes de génération de contenu) seront passibles d'une amende de 2 500 € et de douze années d'emprisonnement

[87] Matthew F. Ferraro, *Deepfake Legislation: A Nationwide Survey—State and Federal Lawmakers Consider Legislation to Regulate Manipulated Media*, 2019, https://www.google.com/url?q=https://www.wilmerhale.com/-/media/a88d81ebebae4ee2aa522242682cabae.pdf&sa=D&ust=1582797385588000&usg=AFQjCNEM2CuvpuGwlRj89q6_8F957FkyPA

Au niveau international on peut également citer les exemples suivants de réglementation :

- Le Kenya, plus drastique, a listé dans une loi de 2019 dix-sept cybercrimes dont la publication de *fake news*, faisant encourir à leurs auteurs deux ans d'emprisonnement et une amende de cinq millions de shillings kényans (environ 42 000 €).
- En Malaisie comme en Indonésie, c'est une peine de six ans d'emprisonnement qu'encourent les auteurs de fausses nouvelles, ce qui, comme l'ont relevé certains commentateurs, laisse craindre de potentielles dérives vers une censure politique.
- La Chine, notamment en réaction aux émotions suscitées par l'application chinoise Zao, a quant à elle adopté tout récemment, le 29 novembre 2019, une loi entrée en vigueur le 1er janvier 2020 qui interdit les *fake news* et impose que les contenus générés par des algorithmes soient identifiés distinctement comme tels. Elle impose par ailleurs aux plateformes de s'assurer de l'identité réelle des utilisateurs et de retirer immédiatement les contenus repérés comme étant illégaux (avec la possibilité de les remettre en ligne en mentionnant l'origine algorithmique du contenu).

Dans d'autres pays[88], des projets de loi sont en cours (Croatie, Corée du Sud) ou alors la question est à l'étude (Belgique et Singapour).

C'est donc au vu des pratiques mises en place dans les autres pays du monde que nous avons formulé les propositions suivantes.

3.7. Nos propositions

Nos propositions sont les suivantes :

- **Développer une organisation supranationale de cyber-coopération policière et judiciaire centrée sur la matière afin de faciliter les échanges d'informations entre les Etats.** Développer une organisation supranationale de cyber-coopération et de

[88] Harold Grand, *En Estonie, une intelligence artificielle va rendre des décisions de justice*, 2019, https://www.lefigaro.fr/secteur/high-tech/en-estonie-une-intelligence-artificielle-va-rendre-des-decisions-de-justice-20190401

cyberentraide policière et judiciaire centrée sur la matière afin de faciliter les échanges d'informations entre les Etats, sur le modèle d'Europol qui existe en matière de lutte contre la criminalité et le terrorisme ou encore d'Iproceeds, mais qui a ce stade voit son objet spécifiquement centré sur l'aspect financier, savoir rechercher, saisir et confisquer les recettes provenant de la cybercriminalité et prévenir le blanchiment d'argent sur l'Internet.

- Permettre dans certains cas exceptionnels (légalement définis), **lorsqu'ils diffusent un contenu clairement illégal, d'ordonner la fermeture définitive d'un site**, d'un profil, d'un compte, d'une page, d'un groupe, sans que le site attaqué puisse se défendre

- **Limiter et quantifier légalement le délai de retrait d'un contenu illicite** signalé aux plateformes sous peine de sanctions financières lourdes

- Obliger les plateformes à **vérifier** les données d'identification lors de la création d'un compte et, en **l'absence d'identification réelle de l'auteur, fermer le compte sans délai**

- Imposer aux hébergeurs situés hors de France de désigner un représentant en France qui assume la responsabilité de l'hébergeur qui n'aurait pas retiré dans ce délai légalement fixé un contenu signalé comme illégal

- Assortir systématiquement les sanctions financières infligées aux plateformes d'une mesure de publicité

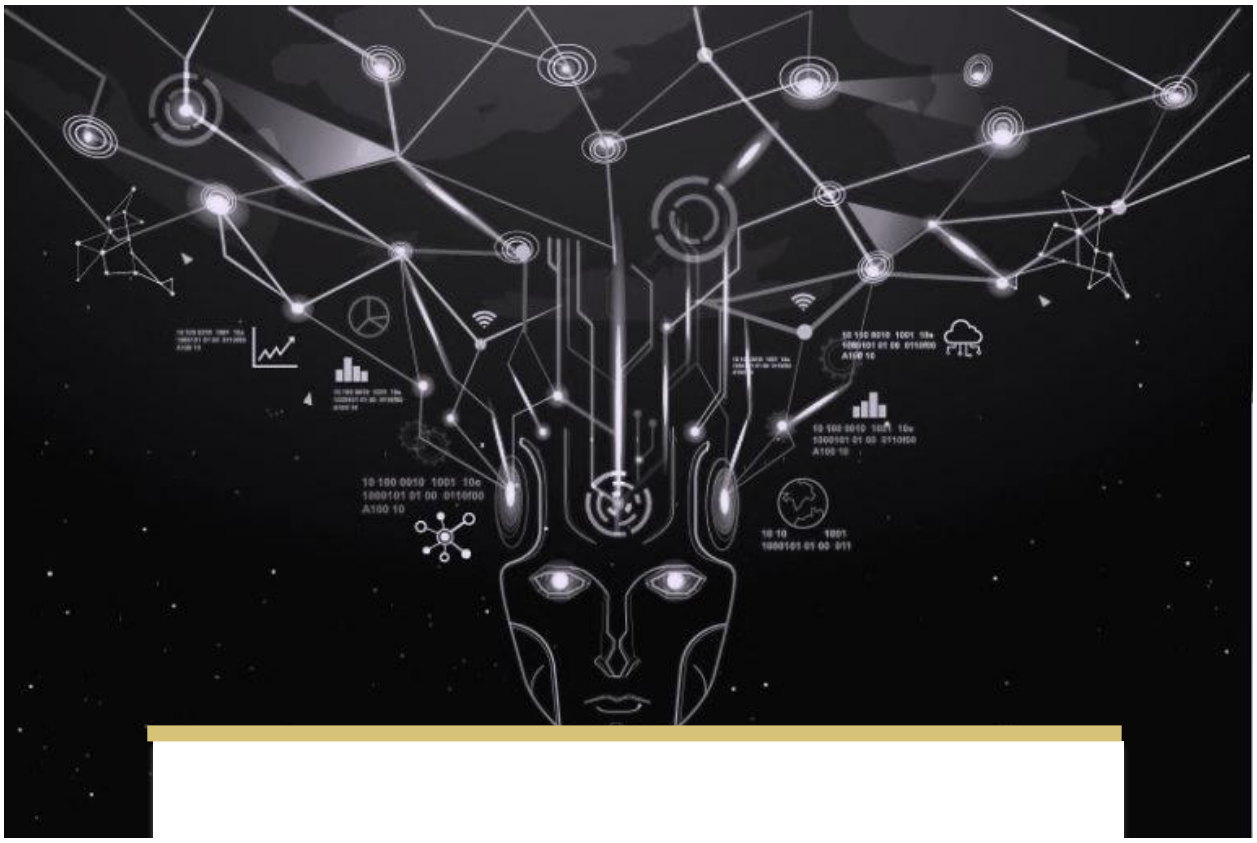
Dans ce cadre, eu égard à la nature transfrontière de l'Internet et des contenus y transitant, une loi nationale si complète et protectrice soit telle ne suffira jamais à permettre de lutter véritablement efficacement contre les deepfakes. C'est la raison pour laquelle, nous suggérons de renforcer au niveau étatique mais surtout supra-étatique (la transmission d'informations ne connaissant pas de frontières) la responsabilité des plateformes.

Proposition n°9

Renforcer au niveau étatique mais surtout supra-étatique (la transmission d'informations ne connaissant pas de frontières) la responsabilité des plateformes.

Dans cette voie :

- **Développer une organisation supranationale de cyber-coopération policière et judiciaire centrée sur la matière afin de faciliter les échanges d'informations entre les Etats.**
- Permettre dans certains cas exceptionnels, **lorsqu'ils diffusent un contenu clairement illégal, d'ordonner la fermeture définitive d'un site**, d'un profil, d'un compte, d'une page, d'un groupe, sans que le site attaqué puisse se défendre
- **Limiter et quantifier légalement le délai de retrait d'un contenu illicite** signalé aux plateformes sous peine de sanctions financières lourdes
- Obliger les plateformes à **vérifier** les données d'identification lors de la création d'un compte et, en **l'absence d'identification réelle de l'auteur, fermer le compte sans délai**
- Imposer aux hébergeurs situés hors de France de désigner un représentant en France qui assume la responsabilité de l'hébergeur qui n'aurait pas retiré dans ce délai légalement fixé un contenu signalé comme illégal
- Assortir systématiquement les sanctions financières infligées aux plateformes d'une mesure de publicité



4. Comment reconnaître un contenu généré par un algorithme ? Comment aller plus loin pour se protéger du *fake content* ?

Malgré des progrès extrêmement rapides, et en dépit de notre volonté d'améliorer notre capacité à détecter un faux, le "faussaire" s'adaptera et sera encore capable de générer un faux encore moins détectable, il existe quelques outils pour accorder plus ou moins de confiance dans une photo, une vidéo, une voix ou un texte.

Nous évoquerons dans un premier temps le moyen de détection le plus élémentaire qui est le signalement par les auteurs de *deepfakes*. Puis, nous dévoilerons quelques techniques ludiques qui permettent, avec un peu d'entraînement, de détecter une image générée par un algorithme. Nous aborderons, ensuite, l'épineux sujet de la détection de faux par des algorithmes. En effet, les algorithmes de génération de contenu créent des faux mais se reposent intrinsèquement sur des détecteurs de faux. Il est donc possible de détecter un faux par un algorithme proche de celui utilisé pour sa création. Nous discuterons, par la suite, de solutions plus accessibles mais nous paraissant néanmoins de plus en plus indispensables pour authentifier et certifier un contenu. Enfin, notre conclusion insistera sur la nécessité de sensibiliser les populations et d'investir dans la recherche scientifique.

Proposition n°1
Soutenir l'amélioration et le développement des techniques de classification de contenu, vrai, faux ou suspect (*fact-checking*, *watermarks*, algorithmes, *blockchain*), tout en définissant des standards européens, afin d'accompagner la population dans l'identification des contenus fiables.

4.1. Détection de faux

4.1.1. Fact-checking

Dans un monde où la fabrication de faux contenu devient de plus en plus facile, une première réponse apparaît dans les années 90s : les sites de *fact-checking*. On note par exemple la création de Snopes[89] en 1995, de *TruthOrFiction.com* en 1999 puis de FactCheck.org[90] en 2003 et de Politifact en 2007. Signalons également l'initiative *www.eufactcheck.eu* de l'EJTA (European Journalism Training Association). Ce projet regroupe une vingtaine d'écoles de journalisme sur le continent européen et propose une méthodologie d'enseignement du fact checking élaborée par les enseignants-journalistes.

Différents journaux ont désormais leur propre blog de *fact-checking* : "les Décodeurs" pour Le Monde, "Scan politique" pour Le Figaro, "Désintox" pour Libération.

Vu la masse d'information, le *fact-checking* ne peut être fait pour toute information par des journalistes d'investigation. En vue d'automatiser ces tâches, huit ans après avoir lancé son blog de *fact-checking* "les Décodeurs", Le Monde a annoncé en février 2017 le lancement de son outil de vérification automatique des sources "Décodex". Décodex propose notamment un plug-in qui permet de vérifier de classifier la source d'une information. Le plug-in vérifie si une source est référencée ou non dans sa base de données et lui attribue une classification.

[89] **Snopes** est un site Web anglophone lancé en 1995 par Barbara et David Mikkelson qui vise à limiter la propagation de fausses nouvelles et de rumeurs infondées qui circulent sur internet. En 2016, Facebook s'est associé avec Snopes pour lutter contre les fake news. Mais en 2019, Snopes s'est retiré de cette collaboration, estimant la charge de ses équipes était trop importante, que l'activité n'était pas assez rémunérée et que son impact n'était pas celui escompté.
[90] **FactCheck.org** est un site internet d'information américain spécialisé dans la vérification des faits, propriété de l'University of Pennsylvania (Annenberg Public Policy Center of the University of Pennsylvania).

Source : Le Monde, Les Décodeurs[91]

Néanmoins nous observons plusieurs limites à ce genre d'outil. D'abord, un site fiable comme Le Monde peut très bien être victime de Fake News et en relayer sans s'en rendre compte. D'autre part il est très compliqué de définir une source comme fiable ou non. Les outils de classification de source sont un outil parmi d'autres pour repérer une fausse nouvelle mais se reposer entièrement dessus pour avoir confiance en un contenu ne paraît pas suffisant.

[91] Lien vers la page : http://www.lemonde.fr/les-decodeurs/article/2017/01/23/l-annuaire-des-sources-du-decodex-mode-d-emploi_5067719_4355770.html

Les journalistes ont besoin de plus d'outils pour les aider à identifier un faux contenu, au-delà de la fiabilité de la source. Dans cette veine, nous voyons apparaître le journalisme augmenté où le journaliste se repose sur de nouveaux outils basés notamment sur l'intelligence artificielle.

Le prototype du Washington Post, *Truth Teller* lancé en 2013 se propose par exemple d'analyser un discours politique en temps réel. Leur outil est basé sur un algorithme de réseaux de neurones profond pour transcrire un dialogue en texte et un autre algorithme qui extrait le fait du texte généré et le compare à d'autres textes.

Jigsaw[92], a annoncé en 2016 le lancement du projet expérimental *ProjectAssembler*[93] en partenariat avec Google et de nombreuses universités. Son objectif est de créer des outils innovants de détection de faux contenu. *"Nous avons construit deux nouveaux détecteurs pour tester sur la plateforme. Le premier est le détecteur StyleGAN pour traiter spécifiquement les deepfakes. Ce détecteur utilise l'apprentissage automatique pour différencier les images de personnes réelles des images deepfake produites par l'algorithme StyleGAN. Notre deuxième modèle, le modèle d'ensemble, est formé à l'aide de signaux combinés de chacun des détecteurs individuels, ce qui lui permet de voir plusieurs types de manipulation simultanément."*[94]. L'outil a été mis en ligne début février 2020 et des tests sont actuellement réalisés par un panel de journalistes.

Ce genre d'outils nous apparaît indispensable à développer pour mieux soutenir les journalistes dans la lutte contre le faux contenu. La seule investigation journalistique ne peut suffire pour vérifier la masse d'information circulant sur le web et notamment les réseaux sociaux comme Twitter et Facebook.

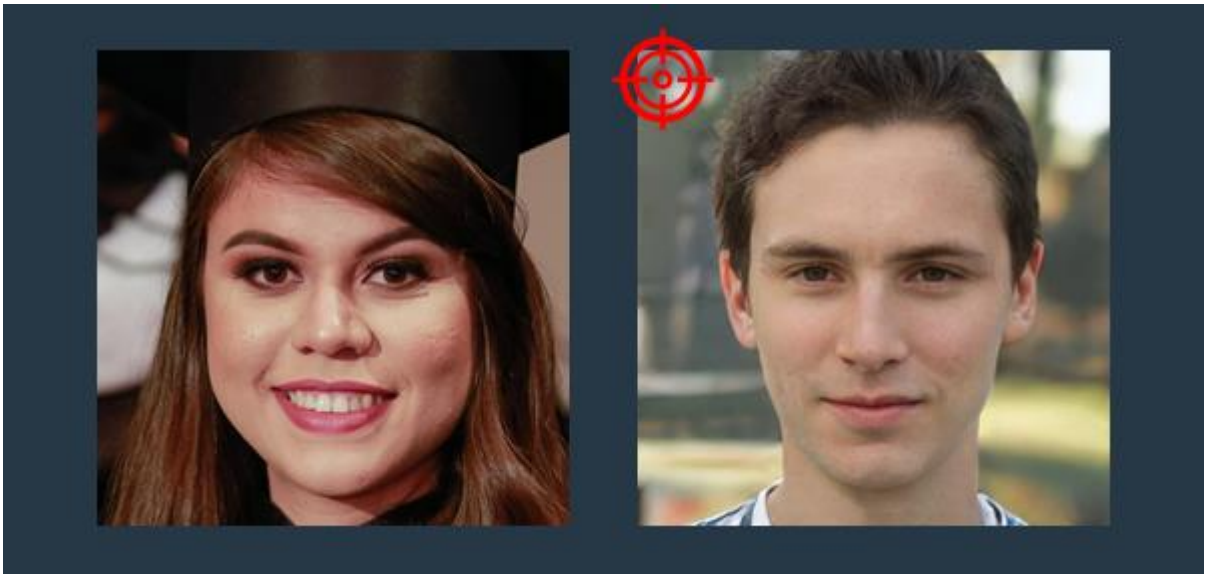
[92] filiale d'Alphabet, anciennement *Google Ideas*
[93] Jay Peters, *Alphabet's Jigsaw unveils a tool to help journalists spot deepfakes and manipulated images*, 2020, https://www.google.com/url?q=https://www.theverge.com/2020/2/4/21122778/alphabet-jigsaw-assembler-tool-news-journalists-deepfakes-images&sa=D&ust=1582797385682000&usg=AFQjCNEbAzrRVVSVtqxnvGSISLsPXlLWwQ
[94] https://medium.com/jigsaw/disinformation-is-more-than-fake-news-7fdd24ee6bf7

4.1.2. Détection visuelle de fausses images, quelques exemples ludiques

Pour créer de parfaites fausses images, les algorithmes ont encore besoin de beaucoup de temps de calcul et de données pour s'entraîner et être suffisamment performants. Par conséquent, dans de nombreux cas, il est encore possible de détecter visuellement des faux à l'aide de quelques défauts récurrents. Au regard de la rapide progression des algorithmes, ces défauts ne seront peut être plus aussi visibles, mais il reste pertinent de connaître quelques indices qui nous permettent d'identifier visuellement de fausses images. Cette section ludique nous paraît toujours très utile pour impliquer la population sur le sujet et la sensibiliser aux enjeux des *deepfakes*.

Dans de nombreux cas, il est encore possible de détecter visuellement des faux à l'aide de quelques défauts récurrents.

Par exemple, lors de la présentation faite par Ekimetrics lors du forum AI Paris en septembre 2020, les intervenants ont demandé à l'audience de sélectionner la photo qui leur semblait fausse parmi une sélection. Voici un premier exemple :



Dans ce premier exemple, la photo de droite est contrefaite alors que celle de gauche est authentique. Nous pouvons le déduire notamment grâce aux indices suivants :

- L'arrière-plan à droite ne semble pas réel.
- Les oreilles de l'individu à droite ne sont pas symétriques et ne semblent pas appartenir à la même personne.

Voici un deuxième exemple :

Ici la photo de gauche est fausse et celle de droite authentique. Nous pouvons le détecter à l'œil nu à l'aide de quelques indices :

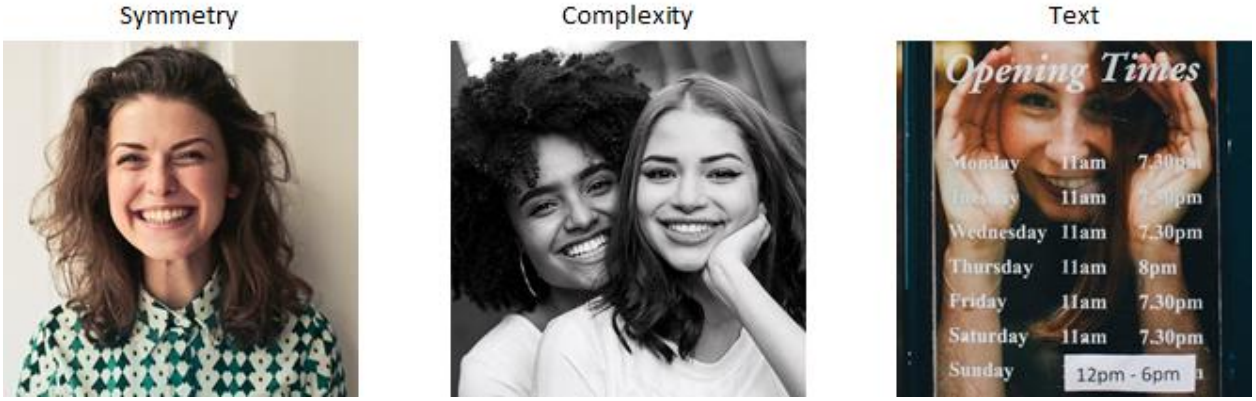
- Les cheveux à gauche ne paraissent pas réalistes, notamment sur la gauche du visage.
- Les habits de la photo de gauche éveillent des soupçons.
- La photo de droite présente un tableau en fond, ce qui est très compliqué à générer par un algorithme.

Il arrive que le rendu soit de qualité très supérieure et ne puisse être détecté au premier coup d'œil :

Source : Jonathan Hui, 2020[95]

Ici, les défauts se situent sur l'asymétrie de l'épaule droite qui n'est pas à la même hauteur que celle de gauche. Le fond de la photo est également suspect.

Bien que de nombreuses fausses photos ne soient pas détectables à l'œil nu, nous avons identifié six signes "classiques" qui permettent d'affirmer qu'un visage est probablement généré par un algorithme : les tâches, l'arrière-plan, les lunettes, l'absence de symétrie, les cheveux et les dents. A contrario, certains signes comme la symétrie, la complexité d'une image ou un texte cohérent sont des éléments qui doivent renforcer notre croyance qu'une photo est vraie.



Les fausses images sont un exemple simple pour introduire le sujet des *deepfakes*. Le pouvoir d'influence des images est tel qu'elles constituent un outil privilégié pour la manipulation d'une audience. En pratique, ces procédés sont utilisés pour créer des applications ludiques telles que

[95] Jonathan Hui, Detect AI-generated Images & Deepfakes (Part 1), *Medium*, 2020

faire vieillir un visage ou le faire sourire. Néanmoins, de faux visages ont été générés pour créer de faux profils sur les réseaux sociaux, comme le profil de la russe Katie Jones évoqué en introduction.

❝

Nous avons identifié six signes "classiques" qui permettent d'affirmer qu'un visage est probablement généré par un algorithme : les tâches, l'arrière-plan, les lunettes, l'absence de symétrie, les cheveux et les dents.

4.1.3. Détection algorithmique

Il existe un algorithme discriminateur dans les *GANs*. Ces modèles sont également capables de détecter certains faux et c'est aussi le cas pour les autres algorithmes de génération de contenu. Trop peu utilisés encore dans cet objectif, il nous semble utile d'investir dans la recherche sur ces applications nécessaires.

Deepfake Detection Challenge

En 2019, Facebook, Amazon, Microsoft, le *Partnership on AI* ainsi que plusieurs instituts académiques (dont le MIT) ont lancé un concours de détection des *deepfakes*, le *Deepfake Detection Challenge* (DFDC)[96]. Celui-ci a pour but de développer des outils de détection de *deepfakes*. Ce défi, qui vise principalement les *deepfakes* vidéos, était organisé sur la plateforme Kaggle et a rassemblé plus de 2000 participants. L'idée générale de cette compétition consiste à construire un modèle capable d'identifier les faux contenus sur un dataset de vidéos ; ces derniers avaient été conçus spécialement à cet effet à travers plusieurs algorithmes de génération de contenu. Après qu'un premier modèle a été retenu avec une précision de 85,26%, il s'est trouvé moins performant sur un dataset *black box*. Ainsi, le modèle gagnant a atteint une précision de 65,15% sur ce dataset. L'objectif de ce défi étant de sensibiliser les chercheurs aux problèmes liés

[96]Facebook AI, *Deepfake Detection Challenge Dataset*, 2020, https://deepfakedetectionchallenge.ai/

aux *deepfakes* et de les inciter à mettre en œuvre des technologies innovantes en vue de détecter toute sorte de manipulation. À la suite de ce défi, le directeur de la technologie au sein de Facebook affirme que la plateforme est en train de construire sa propre technologie pour faire face aux problèmes de faux contenus[97]. Nous détaillons en annexe le *Deepfake Detection Challenge*.

Le directeur de la technologie au sein de Facebook affirme que la plateforme est en train de construire sa propre technologie pour faire face aux problèmes de faux contenus.

Détection algorithmique de faux textes

Plusieurs entités se sont mobilisées pour développer des algorithmes capables de détecter des faux textes générés par des algorithmes. Ainsi, le gouvernement américain — par l'intermédiaire de son agence de recherche pour la défense (DARPA) — a lancé des appels à projets sur le sujet. Notamment, le programme *SemaFor*, ou *Semantic Forensics*, a pour objectif de développer une technologie innovante qui analysera les médias afin de détecter si le contenu est généré ou modifié algorithmiquement (comme par des *GANs*). Le projet intégrera plusieurs types d'algorithmes :

- Les algorithmes de détection sémantique ainsi que les modèles d'attribution, seront capables de trouver si l'origine du *fake* provient d'une entreprise ou d'un particulier identifié.
- Les algorithmes de caractérisation permettront de déterminer si le faux contenu a été généré dans un but malicieux ou non.

[97] James Vincet, Facebook contest reveals deepfake detection is still an 'unsolved problem', *The Verge*, 2020, https://www.theverge.com

L'ensemble de ces technologies aidera le gouvernement à identifier et détecter les campagnes de désinformation[98].

Une approche simple pour détecter le faux texte est la méthode TF-IDF, *Term frequency-inverse document frequency*. Cette méthode permet de déterminer les mots les plus importants en effectuant une pondération de la fréquence relative des termes (*Term Frequency* ou TF) par rapport à la fréquence et distribution de ce même termes dans tous les autres documents (*Inverse document frequency* ou IDF). L'importance du mot augmente proportionnellement en fonction du nombre de fois où il apparaît dans le texte, laquelle est toutefois compensée par la fréquence de ce mot dans le corpus. Ainsi, des mots communs trouvés dans chaque document, notamment les connecteurs logiques, vont appartenir à des rangs inférieurs. En revanche, si le mot *deepfake* apparaît plusieurs fois dans un document sans qu'il ne soit répété dans le corpus, il va être associé à des poids élevés. La méthode TF-IDF ne prenant en compte que des super-pondérations brutes, elle n'est pas capable d'identifier les relations sémantiques. Pour certains cette méthode reste non-optimale, servant à la recherche d'informations plutôt qu'à l'optimisation du contenu[99]. Cependant, la faiblesse de cette méthode réside surtout dans le fait qu'elle nécessite un nombre de mots élevé pour être performante. Ainsi, les résultats sont peu probants sur des textes courts de description de produits, ou même sur des articles de blog.

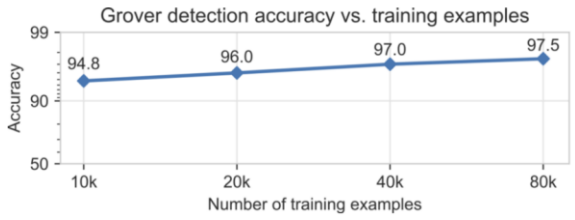
Un autre type d'algorithme de la famille des *Transformers* comme GPT-3, BERT ou *Bidirectional Encoder Representations from Transformers* (Représentations de codeurs bidirectionnels à partir de transformateurs), a été publié par Google AI Language. BERT est doté de plusieurs fonctionnalités, y compris la réponse aux questions et l'inférence en langage naturel. L'innovation principale dans cet algorithme réside dans sa capacité à lire le texte non seulement de gauche à droite, mais aussi de droite à gauche. Ainsi, à l'aide de transformateurs ou *Transformers*, BERT est capable d'apprendre les relations contextuelles entre deux mots, tout en se basant sur les mots qui les entourent. De ce fait, son décodeur est capable de prédire le prochain mot[100]. En masquant

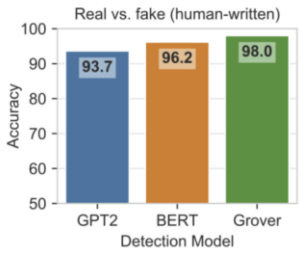
[98] Dr. Matt Turek, Semantic Forensics (SemaFor), DARPA, https://www.darpa.mil/program/semantic-forensics
[99] Stephen Jeske, TF-IDF (Term Frequency-Inverse Document Frequency) Explained, *Learning Hub*, 2019: https://learn.g2.com/tf-idf
[100] Rani Horev, BERT Explained: State of the art language model for NLP, towards data science, 2018

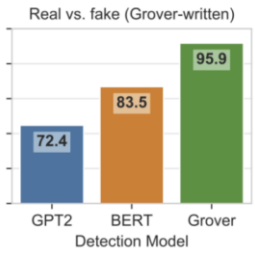
des mots aléatoirement, BERT est capable de prédire des faux contenus. Dans un premier stade de pré-entraînement, deux étapes d'apprentissage non-supervisées sont appliquées. La première, consiste à masquer 15% des mots d'un document pour le prédire par la suite, à travers la méthode *Masked LM,* tout en se basant sur le reste du contenu. Ultérieurement, le modèle procède à la prédiction de la phrase prochaine (NSP ou *Next-Sentence Prediction*), avec des méthodes telles que les inférences en langage naturel et le *Question Answering* (La Réponse à une Question). Dans une deuxième étape, BERT applique le processus de *Self Attention*, ou Auto-attention, pour apprendre les corrélations entre les mots courants et précédents, permettant ainsi "de capter l'attention croisée bidirectionnelle entre des paires de phrases"[101].

Recherche universitaire

Des travaux récents réalisés par un groupe de chercheurs, le *Allen Institute for AI*, ont permis d'élaborer un modèle appelé Grover[102] qui génère et détecte les *fake news* (pour le texte) avec une précision bien supérieure à tous les autres modèles existants (92% contre 73% pour les meilleurs modèles discriminateurs). Dans un rapport daté du 29 octobre 2019, l'équipe de recherche détaille comment se défendre contre les *fake news*[103]. Selon le *Allen Institute for AI*, des générateurs de *fake news* comme Grover sont les plus à même de détecter les *fakes*, étant donné qu'ils détiennent déjà des exemples de faux. Sur 5000 exemples de *fakes news* générées par son réseau de neurones, Grover était capable de différencier les contenus humains de ceux générés par le réseau avec 92% de précision. La capacité de ce modèle à détecter les faux augmente avec le nombre d'exemples fournis.







[101] Sadrach Pierre, Fake News Classification with BERT, *towards data sicence*, 2019: https://towardsdatascience.com/

[102] Allen Institute for AI, *Grover- A State-of-the-Art Defense against Neural Fake News,* https://grover.allenai.org/

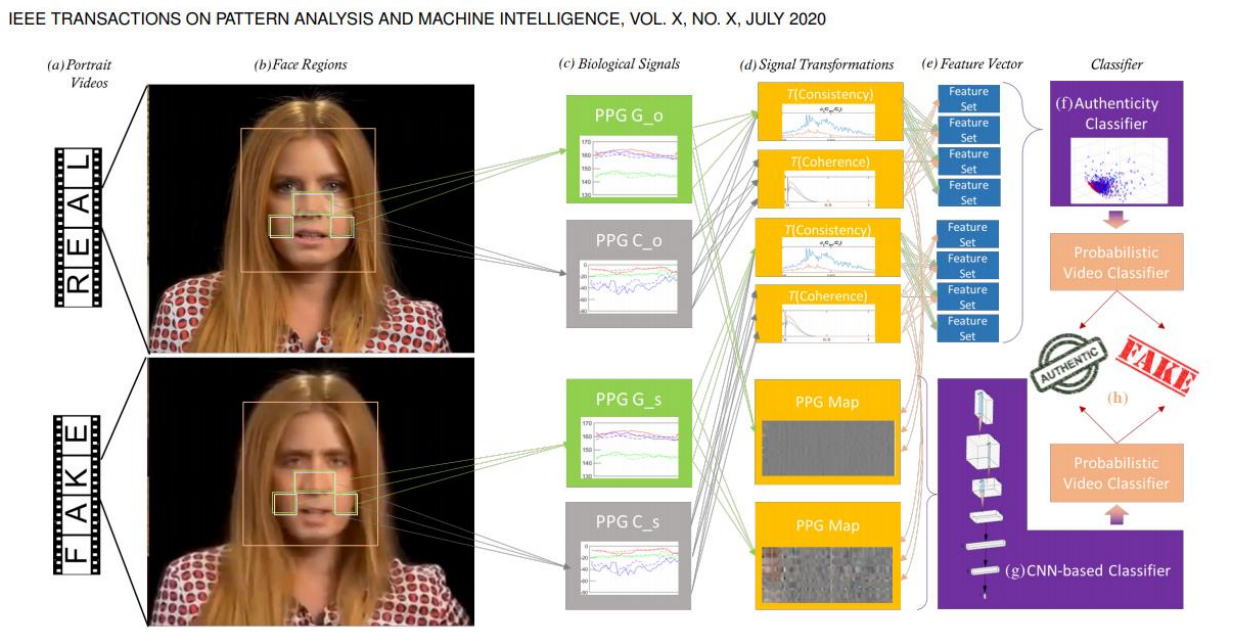
[103] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, Y. Choi & P. G. Allen, *Defending Against Neural Fake News*, 2019, https://arxiv.org/pdf/1905.12616.pdf

Source : Rowan Zellers, *Medium*, 2019

La capacité de Grover ne se restreint pas à la détection des contenus générés par lui-même, car il est aussi capable de signaler les *fakes news* générées par GPT-2 (avant la sortie de GPT-3), avec une précision de 96,1%. Grover est par ailleurs capable de détecter de fausses nouvelles créées par l'humain[104].

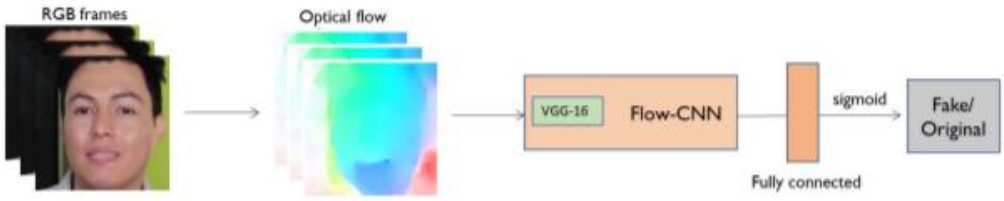
Trois chercheurs, deux de la *Binghamton University* et un d'*Intel*, ont publié récemment un article dans lequel ils décrivent la possibilité de détection des *fakes* dans les vidéos, suite à leur création de *FakeCatcher*, un logiciel capable de signaler les vidéos réelles des personnes à travers les signaux physiologiques non visibles à l'œil nu. Leur affirmation principale vient du fait que les signaux biologiques, qui ne sont pas conservés, ni spatialement ni temporellement dans des faux contenus, peuvent être déployés comme des descripteurs implicites de l'authenticité. Le processus se traduit par la collecte en première étape, des segments vidéo de longueurs fixes, contenant des régions d'intérêts au sein du visage. Par la suite, les signaux biologiques ainsi que leurs transformations et corrélations sont analysés. Puis, les résultats sont utilisés pour créer un classifieur de faux contenus tout en analysant ces transformations et les caractéristiques correspondantes afin de créer une authenticité généralisée. Le classifieur traditionnel est amélioré par la suite à travers la génération de nouvelles cartes de signaux ainsi qu'à travers l'utilisation d'un *Convolutional Neural Network* (CNN), capable de détecter les contenus synthétiques. Comme dernière étape, les probabilités d'authenticité sont agrégées et classifiées permettant de générer une décision binaire distinguant le faux contenu du vrai[105].

[104] Rowan Zellers, Counteracting neural disinformation with Grover, *Medium*, 2019.
[105] Umur Aybars Ciftci et al., FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. X, NO. X, 2020



Source : Umur Aybars Ciftci et al., *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020

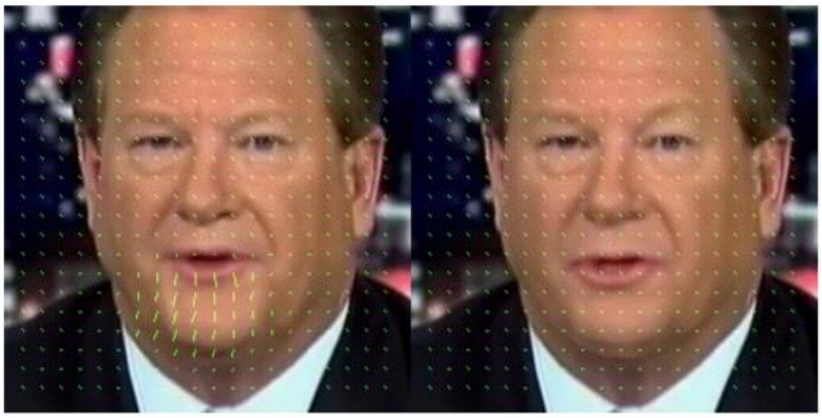
Toujours dans le domaine de la recherche, un article publié en 2019 par le *Media Integration and Communication Center* (MICC) de l'Université de Florence, ainsi que le *National Inter-University Consortium for Telecommunications* (CNIT) de Parme en Italie, exploite la détection des *deepfakes* dans les vidéos à travers le *Optical flow* (flux optique) appliqué aux réseaux de neurones convolutifs[106]. Dans une approche différente de celles qui existent déjà, les chercheurs présentent une approche séquentielle afin d'investiguer l'existence de dissimilarités entre les structures temporelles des vidéos.



Source: Irene Amerini et al., Deepfake Video Detection through Optical Flow based CNN, 2019.

[106] Irene Amerini et al., Deepfake Video Detection through Optical Flow based CNN, 2019.

Par définition, le flux optique est le mouvement apparent sur une scène visuelle, des objets ou des surfaces. Il est provoqué par le mouvement relatif entre la scène elle-même et un observateur (l'œil ou la caméra, par exemple). Ainsi, l'objectif de leur étude est d'utiliser le flux optique pour extraire les disparités entre les images, permettant de discriminer entre celles créées synthétiquement et celles qui appartiennent à de vraies vidéos (mouvements non-usuels dans le visage au niveau des lèvres ou des yeux). De ce fait, le flux optique représentait un vecteur calculé sur deux trames consécutives, afin de retenir les mouvements entre l'observateur et la scène.



Source: Irene Amerini et al., Deepfake Video Detection through Optical Flow based CNN, 2019.

Les résultats des flux optiques calculés sont présentés sur la figure ci-dessus. On est capable d'observer dans un premier temps, les flux sur la vidéo originale (à gauche), et ceux sur le *deepfake* (à droite). Le bruit autour du menton est plus fortement visible dans l'image de la vraie vidéo que celle transformée.

Proposition n°2
Investir dans la recherche scientifique sur les algorithmes de génération de contenu, pour être mieux à même de les combattre mais aussi être capable de les utiliser dans des applications positives pour la société.

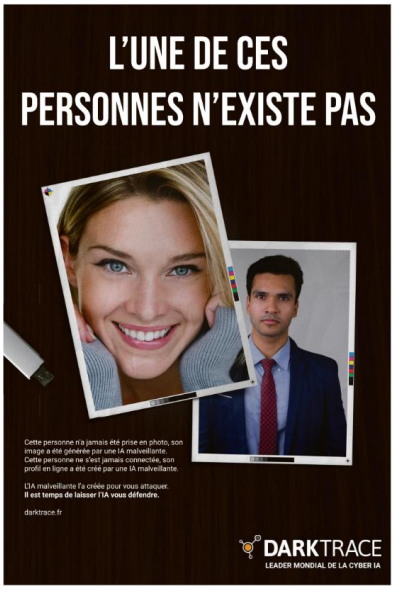
Il faut néanmoins garder en tête que plus les algorithmes évoluent pour permettre de détecter les faux contenus plus ils peuvent être utilisés pour générer du contenu réaliste et indétectable.

De nombreux travaux sont en cours et cette section n'est pas exhaustive.

Proposition n°3
Favoriser les partenariats européens, mutualiser les ressources pour se donner les moyens de rattraper l'avance de la Chine, de la Russie et des Etats-Unis dans ce domaine.

Acteurs du domaine privé

Dans le domaine de la cyberdéfense, la capacité de détecter du faux contenu devient un enjeu majeur. L'entreprise Darktrace, licorne britannique fondée en 2013 spécialisée en cyberdéfense, a lancé récemment une campagne publicitaire dans la presse nationale qui a pour vocation de sensibiliser le grand public aux problématiques d'images générées par l'IA.



Campagne marketing de Darktrace, Septembre 2020

Cette entreprise britannique, valorisée près de 2 milliards de dollars, applique les principes de l'IA au domaine de la cyberdéfense, en développant des algorithmes capables de détecter de nombreuses attaques informatiques telles que les attaques internes, les vulnérabilités latentes, les menaces au développement lent et l'espionnage d'état. Sa technologie principale se base sur les probabilités Bayésiennes et est inspirée du fonctionnement du système immunitaire humain[107].

Un autre exemple, la startup française BusterAI fondée en 2019 par deux français, Julien Mardas et Aurélien Cluzeau. Face à la quantité d'informations critiques diffusée en ligne, ils se proposent de fournir des outils basés sur l'IA pour améliorer la qualité des contenus publiés en ligne. BusterAI développe des algorithmes qui ont vocation à détecter les *fakes* dans les images, les textes et les vidéos pour aider de nombreux médias, par exemple TF1[108]. En fournissant aux entreprises une API, ces dernières sont capables détecter les faux contenus. L'entreprise fournit un tableau de bord ayant comme but de diagnostiquer les contenus, de générer des scores de confiance, d'identifier des biais, ou même d'analyser des sentiments.

4.2. Outils pour protéger le vrai

4.2.1. Filigranes *(Watermarks)*

A défaut de pouvoir parfaitement identifier les contenus erronés, il est nécessaire de mettre en place plusieurs moyens pour authentifier et certifier la véracité d'un contenu. Ici les *watermarks* s'avèrent être un outil indispensable, notamment pour la vidéo et la photographie. Ces domaines sont en effet exposés aux problématiques de piratage mais aussi de modification de contenu. Lorsque des photos et vidéos sont produites, elles peuvent être authentifiées grâce à des marquages visuels (filigranes ou en anglais, *watermarks*[109]) ou digitaux (*metadata*[110]). Le marquage

[107] Darktrace Immune System White Paper https://www.darktrace.com/en/resources/wp-platform.pdf?utm_source=darktrace&utm_medium=products-enterprise

[108] Challenges.fr, Buster AI : l'antivirus de l'information, *Challenges.fr*, 2020

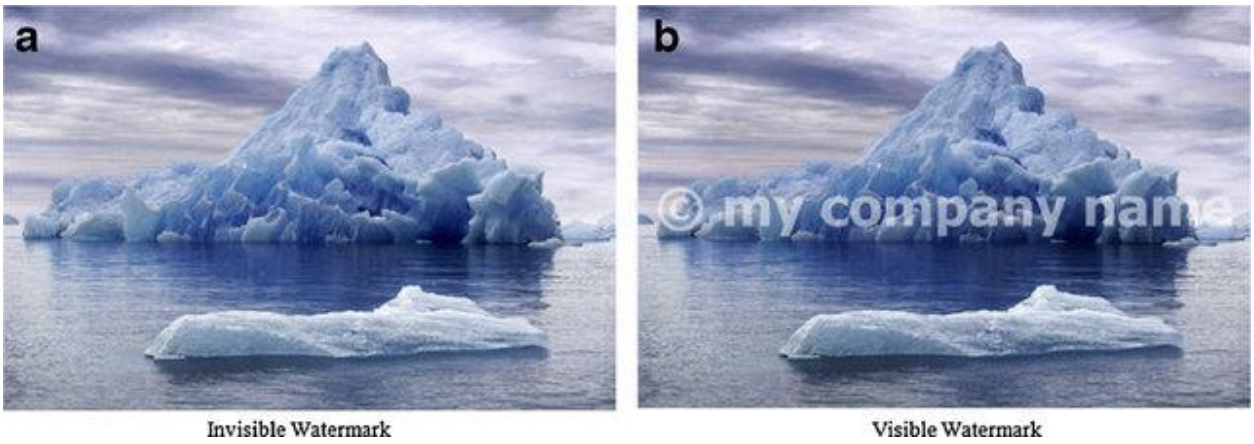
[109] Le **watermark** ou filigrane numérique est une technique permettant d'ajouter des informations de copyright ou d'autres messages de vérification à un fichier ou signal audio, vidéo, une image ou un autre document numérique.

[110] Les **metadata** (ou métadonnées) sont des données qui fournissent de l'information sur une autre donnée. A titre d'exemple pour une image, les métadonnées sont la taille de l'image, sa date de création ou son nom.

des *metadata* peut facilement être modifié par l'utilisation d'un simple logiciel ou une capture d'écran qui permet d'outrepasser les *metadata*. Les photographes peuvent utiliser des sites comme Google Image Search or Digimarc pour effectuer des recherches par image pour contrôler l'utilisation de ces images.

On distingue plusieurs types de *watermarks* :

- *Watermark* visible ou perceptible : modification visible ou perceptible du fichier. Par exemple en ajoutant © ou nom de l'auteur sur une photo.



Exemple de watermark invisible (gauche) et visible (droite)

Source : Sherine M. Youssef et al., 2014[111]

- *Watermark* invisible : il modifie le signal de façon systématique mais imperceptible par l'utilisateur. Cette forme de filigrane peut être considérée comme une forme de stéganographie. Cela peut se faire par l'ajout d'un élément comme un pixel. Il est possible d'ajouter des *watermarks* invisibles dans une photo, vidéo ou même dans un fichier audio.
- *Watermark* fragile : *watermark* invisible qui, une fois le fichier source altéré, devient visible. Cela permet de garantir l'authenticité d'un fichier. Dans le cas d'un fichier audio, Universal Music Group est reconnu pour appliquer des *watermarks* à sa musique[112].

[111] Sherine M. Youssef et al., Adaptive video watermarking integrating a fuzzy wavelet-based human visual system perceptual model, *Multimedia Tools and Applications*, 2014
[112] Hydrogene Audio, 2019, https://hydrogenaud.io/index.php?topic=117462.0

Il existe différentes attaques contre les *watermarks* : l'attaque passive qui ne fait que déceler la présence d'un *watermark* et l'attaque active qui essaie d'éliminer le *watermark*. Les attaques passives sont utilisées en stéganographie alors que l'attaque active est généralement malveillante (tentative d'attaque au copyright, *fingerprint*). Une attaque active peut potentiellement s'effectuer à l'aide d'un algorithme de génération de contenu, qui effacera les *watermarks* de façon systématique. L'attaque active a pour but d'utiliser une image par exemple à l'insu de son auteur, ou permettre d'anonymiser un document.

Ainsi, les *watermarks* ne sont pas infaillibles, qu'ils soient visibles ou non. Les techniques de défenses visibles ont souvent une efficacité limitée car elles altèrent la qualité du fichier. Le *watermark* invisible permet de tracer l'existence d'un fichier et de garantir son intégrité. Il convient à présent de comprendre comment insérer le *watermark* au fichier pour rendre celui-ci le plus sécurisé possible.

Cependant, les *watermarks* ne se limitent pas uniquement aux images. En effet, pour lutter contre le *hacking* à la voix, la société Adobe travaille activement sur le développement d'un *watermark* numérique qui permettra d'identifier si son logiciel de création de voix a été utilisé.

Afin de garantir l'intégrité des fichiers, les auteurs (photographes, vidéastes, etc.) et les sociétés du numérique produisant des logiciels liés à ces écosystèmes (logiciels de retouche photo tel que la suite Adobe) sont amenés à innover afin de permettre davantage de sécurité et de faire respecter les droits d'auteur notamment.

Deux principales techniques d'insertion de marques existent : celle spatiale et l'autre spectrale. La première consiste à intégrer un élément à un endroit précis du fichier. Dans le cas d'une image, le marquage spatial résiste mal aux attaques de type zoom dans l'image ou le recadrage. A contrario, il est également possible de marquer un fichier par un *watermark* spectral. Celui-ci modifie le fichier en profondeur (on rajoute des *patterns* au sein de l'image, sans altérer celle-ci). Il convient dans ce cas précis principalement de modifier de manière infime mais en suivant le même motif l'ensemble du fichier afin "d'encoder" un motif propre. Connaissant ce motif, on peut garantir l'intégrité du fichier en décodant le motif ultérieurement.

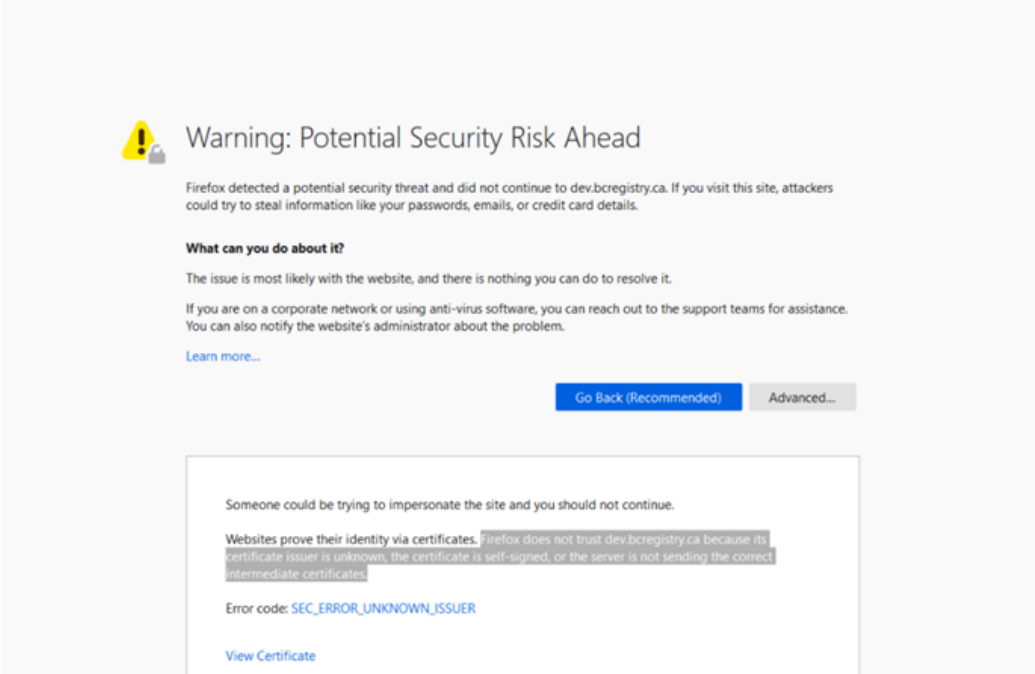
L’insertion d'un marquage à la source permet d’identifier le propriétaire des droits d'une œuvre. L’insertion d'un marquage à l'expédition ou à la lecture permet d’identifier le destinataire ou l'équipement de lecture (*fingerprint*). Cela peut s’avérer très utile pour détecter un utilisateur qui partagerait illégalement le contenu. Cette technique a notamment été utilisée pour marquer des contenus vidéo et identifier les utilisateurs de plateformes payantes qui partageaient les contenus sur internet (*streaming*). C’est par exemple le cas pour les retransmissions en direct par HBO des matchs de boxe.

4.2.2. Signatures numériques

Les *watermarks* sont utiles pour tracer un contenu, mais un autre outil bien connu peut s’avérer très utile pour authentifier un auteur : les signatures numériques, ou signatures électroniques avancées selon les termes de la législation européenne.

La signature numérique est un code mathématique unique généré à partir d’un document et d’une clé (dite privée ou secrète). Une clé jumelle (appelée clé publique) permettra à tous ceux qui en ont connaissance de vérifier l’intégrité du document et l’identité de l’auteur. Les clés publiques peuvent être rassemblées pour former un réseau de confiance (*Web Of Trust*), public ou privé, qui constitue alors un tiers garant des identités.

Cette solution est largement utilisée en informatique et sur le web afin de certifier l’origine d’un site. Par exemple, l’affichage de ce message indique que le site visité n’est pas fiable car son origine n’a pas pu être vérifiée.



C'est un mécanisme simple mais efficace, imposé en France par le règlement eIDAS[113] depuis 2014 et qui permet de proposer des certificats numériques pour personnes physiques.

Il existe trois niveaux de certificat : simple, avancée ou qualifiée ayant une valeur juridique croissante, en fonction de l'usage. Une signature simple pour signer une note de frais jusqu'à une signature qualifiée pour un acte authentique chez le notaire.

La signature électronique peut donc être utile pour s'assurer de l'identité de l'émetteur d'un contenu.

4.2.3. Hachage

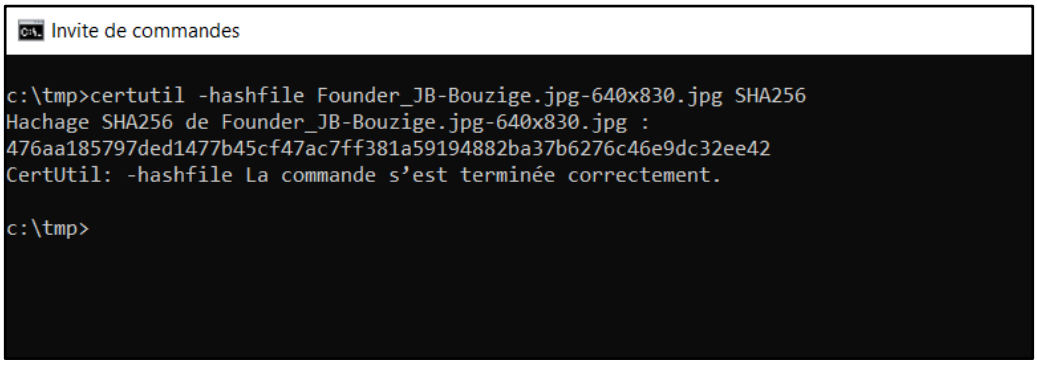
Une fonction de hachage (de l'anglais Hash) permet de calculer une empreinte numérique aussi appelée somme de contrôle, servant à identifier une donnée. Les fonctions de hachage sont utilisées en informatique pour authentifier un mot de passe ou un fichier.

[113] https://www.ssi.gouv.fr/entreprise/reglementation/confiance-numerique/le-reglement-eidas/

Ainsi, en stockant l'empreinte numérique d'un mot de passe cela permet d'éviter de stocker le mot de passe en clair dans une base de données, pour en garantir sa confidentialité, tout en permettant d'identifier celui-ci par comparaison de l'empreintes stockées dans la base de données avec l'empreinte calculée sur le mot de passe saisie par l'utilisateur lors de l'authentification ultérieure.

De même lors d'un transfert de fichier, le calcul de l'empreinte numérique du fichier avant l'envoi puis le calcul de l'empreinte après réception du fichier permet, par comparaison de s'assurer de l'authenticité et qu'il n'a pas été altéré pendant le transport. Cette comparaison est nécessaire notamment en cas d'erreur de transport mais aussi pour se prémunir d'un piratage.

Un exemple de calcul de l'empreinte numérique d'une photo disponible sur le site internet Ekimetrics à partir de l'utilitaire CertUtil, disponible dans Windows 10 :



Dans cette exemple, l'empreinte est calculée en utilisant un algorithme de cryptages SHA-256, considéré comme sûr au sens cryptographique :

"476aa185797ded1477b45cf47ac7ff381a59194882ba37b6276c46e9dc32ee42"

Le hachage permet donc de s'assurer de l'authenticité d'un fichier grâce à la vérification de son intégrité entre deux étapes.

4.2.4. Blockchain

D'autres solutions innovantes comme la *blockchain* émergent pour garantir l'intégrité des fichiers.

Comme détaillé dans le rapport Praxis sur les données médicales[114], la technologie *blockchain* consiste en un réseau de registres de données distribués (aussi appelés blocs) auprès de chaque participant. Les blocs sont liés chronologiquement les uns aux autres. Publique, privée ou hybride, la *blockchain* accepte de nouvelles données dans son registre grâce à un mécanisme de consensus, comme le minage. Associé à un mécanisme de signature numérique, ce consensus pourra s'accorder auparavant sur l'identité de l'auteur[115].

La *blockchain* a de nombreux avantages théoriques :

- **La décentralisation** : les registres de données sont historiquement centralisés par un intermédiaire de confiance (une banque, un hôpital, un ministère, etc.) qui assure la sécurité et l'utilisation des données. La *blockchain* et les registres distribués sur le réseau permettent d'assurer la confiance nécessaire à un système d'échange de données.
- **La sécurité** : dans une *blockchain*, les registres de données ne peuvent être altérés sans le consensus du réseau. C'est la principale force du système qui, à l'opposé d'un registre centralisé sur un serveur, ne peut être hacké. En revanche, il est possible que le mot de passe d'un utilisateur, donnant accès à sa clé privée, soit subtilisé par des moyens frauduleux[116]. L'identité de la victime pourra alors être usurpée.
- **La portabilité** : l'authenticité des données d'un participant peut être vérifiée par plusieurs acteurs si celui-ci les autorise. La possibilité d'enregistrer les données sur une *blockchain* permettrait à plusieurs entités dont les bases de données sont contradictoires (hôpital, pharmacie, etc.) d'assurer une meilleure traçabilité d'un patient.

---

[114] Club Praxis, *Le tsunami numérique peut-il révolutionner nos systèmes de santé publique?*, 2018, https://espas.secure.europarl.europa.eu/orbis/sites/default/files/generated/document/en/Donn%C3%A9es%20m%C3%A9dicales.pdf

[115] Healthcare Data Institute, INTERNATIONAL THINK TANK DEDICATED TO BIG DATA IN HEALTHCARE, 2017, https://healthcaredatainstitute.com/wp-content/uploads/2015/02/presentation-hdi_04122017.pdf

[116] Jeff John Roberts, *Bitcoin Theft: 5 Common Threats, Hacks, and Scams*, 2017, http://fortune.com/2017/12/08/bitcoin-theft/

L'entreprise de sécurité informatique Amber[117], et son Directeur Général Shamir Allibhai, ont mis au point une technique permettant de garantir l'intégrité des vidéos de caméras de surveillance et des caméras-piétons (utilisées par la police) grâce à la *blockchain*. Cela fait suite aux travaux de Josh Mitchell, chercheur en sécurité, qui a prouvé qu'il existe des failles de sécurité graves dans la majorité des caméras utilisées par la police américaine. Le système s'appuie sur la blockchain, en authentifiant des fragments de vidéo : la plateforme génère des "hash" (une signature mathématique unique qui représente le fragment de vidéo) à intervalles réguliers et les partage à une *blockchain* publique. En relisant la vidéo et en la passant au travers du même algorithme, les hash seront identiques (et donc comparables à ceux enregistrés sur la *blockchain*). En revanche, si la vidéo a été modifiée, la signature sera différente, ce qui protège donc de la manipulation de la vidéo elle-même.

L'entreprise de sécurité informatique Amber (…) [mentionne avoir] mis au point une technique permettant de garantir l'intégrité des vidéos de caméras de surveillance et des caméras-piétons (utilisées par la police) grâce à la blockchain.

La *blockchain* est une technologie de stockage et de transmission d'informations, transparente, sécurisée, qui fonctionne de manière décentralisée (sans organe central de contrôle). Par extension, une *blockchain* constitue un historique de tous les échanges effectués entre ses utilisateurs depuis sa création. Cet historique est sécurisé et distribué : il est partagé par ses différents utilisateurs, sans intermédiaire, ce qui permet à chacun de vérifier la validité de la chaîne. Comme l'écrit le mathématicien Jean-Paul Delahaye, il faut s'imaginer "*un très grand cahier, que tout le monde peut lire librement et gratuitement, sur lequel tout le monde peut écrire, mais qui est*

[117] Lily Hay Newman, *A New Tool Protects Videos From Deepfakes and Tampering*, 2019, https://www.wired.com/story/amber-authenticate-video-validation-blockchain-tampering-deepfakes/

impossible à effacer et indestructible". Nous savons néanmoins qu'il existe plusieurs exemples démontrant qu'il n'est pas impossible de reconstituer tout le cahier avec une immense puissance de calcul.

La *blockchain* est une alternative décentralisée d'authentification de contenu, qui agit comme un registre connecté au réseau et sécurisé par la puissance de calcul de celui-ci. Tout comme le *watermark*, son utilisation ne permettrait pas de détecter du faux contenu, mais de garantir l'intégrité de celui-ci. Les fichiers (et leur signature numérique) insérés à la *blockchain* ont un identifiant unique. La *blockchain* est très utile pour authentifier une modification ou transaction sans nécessairement en connaître l'auteur.

Les attaques informatiques contre la *blockchain* sont très difficiles à mener car un acteur ou un groupe doit contrôler plus de 50% des garants qui composent la *blockchain* afin de corrompre son contenu. La *blockchain* agirait comme tiers de confiance décentralisé, et permettrait d'assurer l'intégrité d'un contenu.

La mise en place d'un système d'authentification suppose de construire un marqueur d'image robuste aux manipulations de mise en page des images, et donc repose a priori sur un système de type *watermark*.

Il nous semble nécessaire de tirer profit de ces avancées technologiques pour protéger les citoyens. La recherche scientifique est extrêmement prolifique sur ces sujets essentiels et il convient d'appréhender au mieux ces outils technologiques que sont les *watermarks,* les signatures électroniques et la *blockchain* afin de garantir l'intégrité du contenu en ligne.

4.3. Sensibilisation et éducation

Dans la mesure où les différentes législations, ainsi que les politiques mises en place par les plateformes, sont des garde-fous mais ne seront jamais suffisantes pour se prémunir des effets néfastes de la diffusion des *fake news* ou des *deepfakes*, le développement de l'esprit critique et la sensibilisation des populations à ces thématiques est selon nous un enjeu majeur.

Nous suggérons de mettre en place une campagne grand public de sensibilisation contre les deepfakes. Par exemple en mettant en avant, dans des documentaires ou des formations, de simples astuces de détection de fausses images lorsque le faux est détectable visuellement. Même si de nombreux *deepfakes* ne sont plus détectables à l'œil nu, nombreux sont encore les deepfakes de mauvaise qualité. Ce type de formation permet de sensibiliser la population au sujet des *deepfakes* de manière ludique. Il nous semble essentiel d'enseigner des méthodes qui permettront de se protéger vis-à-vis de ces contenus malveillants, et de faire prendre conscience de la portée que peuvent avoir ces fausses informations quand elles sont largement relayées.

Proposition n°10
Faire de la lutte contre les *deepfakes* un objectif européen pour 2021. Mettre en place des campagnes de sensibilisation pour expliquer de manière pédagogique les *deepfakes*, leurs utilisations et les risques associés en termes de détournement des images, de la voix et des vidéos, et insister sur les applications possibles en manipulation d'opinion.

Il est aussi possible d'aller plus loin et d'éduquer la population sur le fonctionnement des algorithmes. Jean-Baptiste Bouzige, fondateur et Directeur Général d'Ekimetrics, estime que la mauvaise compréhension des algorithmes et des enjeux posés par ces nouveaux outils est au cœur du problème. Il est nécessaire de construire une expertise européenne sur le sujet, et de s'assurer que nous formons des chercheurs qui seront à la pointe sur ce sujet. En France, on note le succès de structures de formation de code comme l'entreprise Le Wagon ou encore l'École 42 fondée par Xavier Niel.

De nombreux outils commencent à se développer côté éducation. On peut citer le langage de programmation *Scratch*, développé par le MIT, qui permet notamment aux enfants de comprendre comment fonctionne un algorithme et de "jouer" avec. Il nous semble important de considérer ces outils pour éduquer les nouvelles générations.

Néanmoins, les *deepfakes* sont pour partie indétectables et tout le monde peut malheureusement en être victime y compris ceux qui sont sensibilisés à ce sujet. Il est donc indispensable d'investir également dans des technologies et des techniques plus poussées, capables de détecter même les *deepfakes* les plus avancés qu'il n'est pas possible de détecter autrement.

4.4. Rôle des auteurs

Le moyen le plus élémentaire pour un utilisateur de savoir si le contenu qu'il a sous les yeux est authentique ou s'il a été manipulé à l'aide d'un algorithme serait tout simplement que l'auteur du contenu signale si c'est un deepfake le cas échéant. Ce signalement peut être volontaire ou contraint, que ce soit par les conditions générales d'utilisation des plateformes ou par la réglementation.

Ainsi, la Chine est, à l'heure où nous écrivons ces lignes, l'un des pays avec la législation la plus stricte en matière de protection contre les deepfakes. En effet, une loi en vigueur depuis le 1er janvier 2020 stipule que "*la publication de contenus vidéo mensongers du genre "deepfake" créés à partir d'une intelligence artificielle doit être signalée par celui qui les a publiés. [...] Avec cette réglementation, les autorités peuvent poursuivre en justice non seulement les simples utilisateurs, mais aussi les services d'hébergement d'images et de vidéos s'ils omettent de signaler les contenus deepfakes*." Cette loi suppose cependant de disposer des technologies de détection des *deepfakes*, afin de pouvoir sanctionner un manquement à l'obligation de signalement.

Le cas de la Chine semble pour le moment isolé, et aucun autre pays à notre connaissance n'impose une obligation de signalement aux auteurs de *deepfakes*. Ce n'est pas non plus le cas des plateformes internet qui adoptent d'autres approches. Comme évoqué plus haut, Facebook supprime les contenus qu'il détecte comme étant du *deepfake* (à l'exception des vidéos ouvertement humoristiques) tandis que Twitter veut lui-même signaler les contenus *deepfakes* qu'il détecte.

Néanmoins cette posture suppose d'avoir des moyens techniques fiables pour identifier les contenus manipulés à l'aide d'un algorithme. En attendant le déploiement de ces moyens

techniques (certains existent déjà), les utilisateurs peuvent se protéger des contenus *deepfakes* par eux-mêmes, car certains détails peuvent en effet trahir la manipulation d'un contenu.

Proposition n°11
Renforcer la responsabilité encourue par les auteurs. Imposer à chaque auteur de signaler tout contenu *deepfake*.

4.5. Rôle des plateformes

La source de documents *deepfakes* peut être difficile à localiser et il peut être difficile pour les autorités de responsabiliser et pénaliser les producteurs de *deepfake*.

Il est important pour ce faire que les plateformes coopèrent avec les autorités et mettent à leur disposition les informations dont elles disposent dans le cadre d'enquêtes ou de procédures judiciaires.

Proposition n°6
Imposer aux plateformes de partager avec les autorités judiciaires les informations qu'elles détiennent sur les auteurs présumés de *deepfakes*, dans le cadre d'enquêtes pénales, sans que les plateformes ne puissent leur opposer de motif légitime de refus.

Les consommateurs de *deepfake* participent, souvent malgré eux, à la redistribution de ces documents. Sans signalements explicites indiquant que ces documents sont des *deepfakes*, les redistributeurs ne sont pas nécessairement responsables. Les plateformes ont donc une part de responsabilité et un rôle important à jouer pour limiter les effets néfastes des *deepfakes*.

Proposition n°4
Renforcer la responsabilité encourue par les plateformes (par exemple au sein du Digital Services Act). Imposer à chaque plateforme de mettre en place une structure de gouvernance interne sur la régulation de leur contenu. Mettre chaque organe de gouvernance sous supervision européenne en s'inspirant de la régulation mise en place dans le domaine bancaire suite à la crise de 2008.

Les plateformes sont capables de mettre en place plusieurs outils de lutte contre les *deepfakes* :

- Une pénalité pour les auteurs publiant du contenu faux ou malveillant de manière non explicite (exemples de pénalité : limitation des publications, limitation temporaire de l'audience de leurs publications, mise en place d'un processus de validation pour ces auteurs suspects, etc.).
- Une mention claire mettant en garde l'utilisateur que la publication n'a pu être vérifiée, voire même qu'elle est suspecte.
- Une limitation du nombre de destinataires d'une publication.

Par exemple, WhatsApp limite déjà le nombre de personnes à qui il est possible de transférer un message, ainsi que la taille des groupes. En France, la taille des groupes est limitée à 256 personnes. Depuis Janvier 2019, pour lutter contre le fausses nouvelles, Whatsapp a limité le transfert de messages à cinq personnes (au lieu de 20). Whatsapp précise également si le message a été transféré, pour permettre aux utilisateurs de mieux transférer ces messages viraux.

L'application Telegram, de son côté met en avant la mise en place un *slow mode* qui permet d'empêcher les utilisateurs de partager plus d'un contenu par unité de temps (par exemple pendant une minute). Néanmoins c'est un mode facultatif, choisi ou non par l'administrateur du groupe. Par ailleurs, la taille d'un groupe Telegram peut atteindre 200 000 personnes.

Proposition n°5

Inciter les plateformes à mettre en place des outils pour diminuer l'impact des *deepfakes*.

- Une pénalité pour les auteurs publiant du contenu faux ou malveillant de manière non explicite.
- Une mention claire mettant en garde l'utilisateur que la publication n'a pas pu être vérifiée, voire même qu'elle est suspecte.
- Une limitation du nombre de destinataires d'une publication.
- Une limitation des transferts d'une publication par utilisateur.

Fraud
High potential to deceive
Fake Is
at Your Door
TRAP
A.I.
BUILDING YOUR STRATEGY
Get a Grip
MOBILIZE
technology
DECODING
Game On!
FEAR...
UNDER INFLUENCE
L'ombre d'un doute
What's up ?
My blindness
Post truth
WARNING:
ASSAUT SUR L'UNIVERS
KNOW YOUR ENEMY
THE POWER OF SUGGESTION
Hoaxes
INSPIRATION IN THE FACE OF ADVERSITY
SHOW ME THE IMPACT

Conclusion

A l'heure où, en France, les réseaux sociaux sont la première source d'information pour les moins de 34 ans, où en Birmanie l'application Facebook est, selon le New York Times, préinstallée sur tous les téléphones et accusée d'incitation à la haine contre les Rohingyas**, l'impact de faux contenus est plus que jamais d'actualité**. D'autant plus que la création de ces deepfakes, créés par les algorithmes de deep learning, des algorithmes d'intelligence artificielle s'inspirant du fonctionnement du cerveau est devenue à la portée de tous et que **leur diffusion est bien souvent virale**.

L'objectif de notre rapport était de trois ordres : **pédagogique** afin de réveiller les consciences sur l'enjeu que représente le développement non contrôlé des algorithmes de générations de contenu ; **politique** en pesant sur le débat autour du cadre numérique, réglementaire et légal des deepfakes ainsi que sur le débat économique et social au niveau français et européen ; **scientifique** en permettant à l'Europe de se positionner comme leader dans le combat contre les deepfakes.

Les algorithmes de génération de contenu, comme toute nouvelle avancée technologique, **cristallisent de nombreuses peurs**. Pour autant, il ne s'agit pas de bannir la technologie à l'origine des deepfakes, non seulement car il n'est à la portée d'aucun gouvernement d'entraver durablement la marche de la science, mais aussi **car ces techniques peuvent avoir des applications utiles dans de vastes domaines** qui vont, comme nous l'avons vu, de l'amélioration des prédictions épidémiologiques à l'assistance au handicap en passant par le divertissement et la création artistique.

Annexes

Annexe n°1 : *Deepfake Detection Challenge* (DFDC)

La génération de contenu, notamment celle des *deepfakes*, est passé d'un statut dit "sophistiqué", à un niveau où toute personne est capable d'en générer.

En 2019, cette initiative collaborative entre plusieurs leaders dans le monde de la technologie ainsi que dans le monde académique, a été lancé autour du sujet de *deepfakes*. Ainsi, le *Deepfake Detection Challenge* (DFDC), regroupe AWS, Facebook, Microsoft et plusieurs autres instituts du milieu académique dont le MIT, avec pour objectif d'inciter les chercheurs à développer de nouvelles technologies qui permettent de remédier aux problèmes des *deepfakes*.

- **Construction du jeu de données**

Ainsi, un *dataset* a été construit de plus de 100.000 vidéos, sur lesquels 8 algorithmes de modification du visage ont été appliqués, tels que les *GANs* et des techniques de raffinement, notamment les superpositions et les inclusions des distracteurs.

Attirant plus de 2000 individus, tous les participants devaient entraîner et tester leurs modèles sur un même *dataset*. Par la suite, ces modèles seront testés sur un *dataset* de *black box*, constitué de vidéos inédites et non partagées avec les participants à l'avance. Des vidéos de tutoriel de maquillage ou de peintures ont été utilisées pour ajouter une complexité aux modèles détecteurs. Des augmentations ont également été appliquées à l'ensemble du *dataset* du *black box*, mais sur l'ensemble de ces tests, des techniques supplémentaires ont été appliquées pour augmenter le niveau de difficulté.

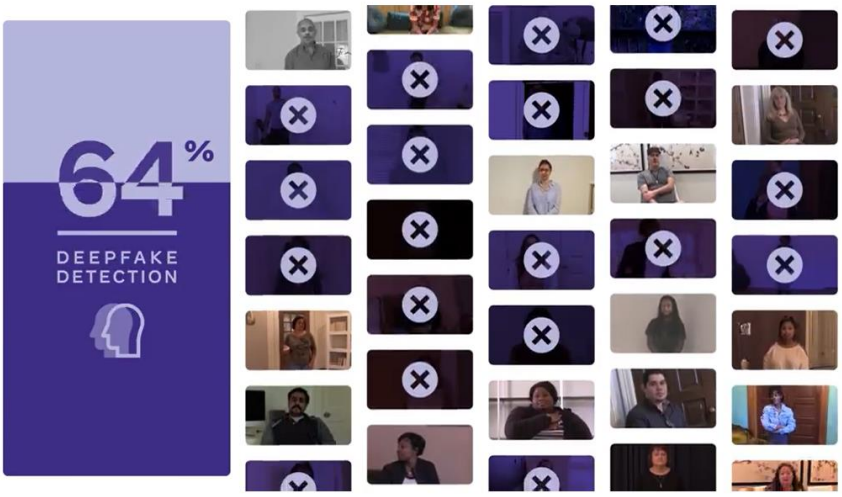
Source: Devin Coldewey, *The Crunch*, 2020 (Credits: Facebook)

Il s'avère parfois difficile, de pouvoir détecter instantanément à l'œil nu, du faux contenu. Prenons l'exemple de ces 6 vidéos générées pour cette compétition. Seulement les images 1, 4 et 6, représentent du vrai contenu.

Pour M. Scroepfer, le directeur de technologie de Facebook, la variance dans le dataset généré pour ce défi, joue un rôle primordial : "*Si votre ensemble de données ne présente pas la variance appropriée dans l'apparence des personnes réelles, votre modèle n'en aura pas une compréhension représentative. Je pense que nous nous sommes donnés beaucoup de mal pour nous assurer que cet ensemble de données soit suffisamment représentatif*". Pour mettre l'accent sur l'importance de la représentativité, M. Scroepfer ajoute que plusieurs dimensions, tel que l'âge, l'origine éthique et le sexe, ont été pris en considération. Pour lui, la technologie développée doit être en mesure de représenter tout le monde.

- **Résultats de la compétition**

Ainsi, le premier modèle qui a été retenu avec une précision de 85,26%, a échoué une fois qu'il était testé sur le *dataset* de *black box*. Cela montre en effet que la détection de faux en nature est un problème très difficile. Malgré des résultats prometteurs, Facebook affirme que la détection des *deepfakes* reste un problème non-résolu.



Le modèle plus performant atteint une précision de 65,15%.

Source : Facebook AI

Après avoir annoncé sa satisfaction, M. Scroepfer a déclaré que Facebook était en train de développer sa propre technologie pour faire face aux problèmes du *fake*. Néanmoins, les algorithmes gagnants de cette compétition seront publiés en open-source afin d'aider d'autres chercheurs. En revanche, Facebook a déclaré qu'il allait garder sa propre technologie de détection secrète pour éviter qu'elle ne soit modifiée par ingénierie inverse.

Annexe n°2 : Les sept principes éthiques de la Commission Européenne en matière d'intelligence artificielle[118]

Les "Lignes directrices éthiques pour une intelligence artificielle digne de confiance" sont un document préparé par le Groupe d'experts de haut niveau sur l'intelligence artificielle (AI HLEG). Ce groupe d'experts indépendants a été mis en place par la Commission européenne en juin 2018.

Fondées sur les droits fondamentaux et les principes éthiques, les lignes directrices énumèrent sept exigences clés auxquelles les systèmes d'IA doivent répondre pour être fiables. Ces exigences s'appliquent aux différentes parties prenantes qui participent au cycle de vie des systèmes d'IA :

- **Les développeurs**, c'est-à-dire ceux qui recherchent, conçoivent et/ou développent des systèmes d'IA ;
- **Les "déployants"**, qui sont les organisations publiques ou privées qui utilisent des systèmes d'IA dans le cadre de leurs activités et pour offrir des produits et des services à d'autres personnes ;
- **Les utilisateurs finaux**, qui s'engagent dans le système d'IA, directement ou indirectement ;
- **La société** dans son ensemble, qui englobe toutes les autres personnes qui sont directement ou indirectement concernées par les systèmes d'IA.

Ces différents groupes de parties prenantes ont des rôles différents à jouer pour garantir le respect des exigences :

- Les développeurs doivent mettre en œuvre et appliquer les exigences aux processus de conception et de développement ;
- Les "déployants " doivent s'assurer que les systèmes qu'elles utilisent et les produits et services qu'elles offrent répondent aux exigences ;

[118] Plus d'information en suivant ce lien :
https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1#Human%20agency

- Les utilisateurs finaux et la société au sens large devraient être informés de ces exigences et pouvoir demander qu'elles soient respectées.

Ci-dessous une explication plus détaillée des sept principes clés énoncés par la commission européenne.

Action humaine et contrôle humain

Les systèmes d'IA doivent soutenir l'autonomie humaine et la prise de décision, comme le prescrit le principe du respect de l'autonomie humaine. Cela implique que les systèmes d'IA doivent à la fois agir comme des facilitateurs d'une société démocratique, florissante et équitable en soutenant l'action de l'utilisateur et en favorisant les droits fondamentaux, et permettre un contrôle humain.

Robustesse, technique et sécurité

La robustesse technique, qui est étroitement liée au principe de prévention des dommages, est un élément essentiel pour parvenir à une IA digne de confiance. La robustesse technique exige que les systèmes d'IA soient développés avec une approche préventive des risques et de telle manière qu'ils se comportent de manière fiable et prévisible tout en minimisant les dommages involontaires et inattendus, et en prévenant les dommages inacceptables. Cela doit également s'appliquer aux changements potentiels de leur environnement opérationnel ou à la présence d'autres agents (humains et artificiels) qui peuvent interagir avec le système de manière antagoniste. En outre, l'intégrité physique et mentale des êtres humains doit être garantie.

Respect de la vie privée et gouvernance des données

Le droit à la vie privée, étroitement lié au principe de prévention des dommages, est un droit fondamental particulièrement affecté par les systèmes d'IA. La prévention des atteintes à la vie privée nécessite également une gouvernance des données adéquate qui couvre la qualité et

l'intégrité des données utilisées, leur pertinence au regard du domaine dans lequel les systèmes d'IA seront déployés, leurs protocoles d'accès et la capacité de traiter les données de manière à protéger la vie privée.

Transparence

Cette exigence est étroitement liée au principe d'explicabilité et englobe la transparence des éléments pertinents pour un système d'IA : les données, le système et les modèles commerciaux.

Diversité, non-discrimination et équité

Afin de parvenir à une IA digne de confiance, nous devons permettre l'inclusion et la diversité tout au long du cycle de vie du système d'IA. Outre la prise en compte et la participation de toutes les parties prenantes concernées tout au long du processus, cela implique également de garantir l'égalité d'accès par des processus de conception inclusifs ainsi que l'égalité de traitement. Cette exigence est étroitement liée au principe d'équité.

Bien-être sociétal et environnemental

Conformément aux principes d'équité et de prévention des dommages, la société au sens large, les autres êtres sensibles et l'environnement doivent également être considérés comme des parties prenantes tout au long du cycle de vie du système d'IA. La durabilité et la responsabilité écologique des systèmes d'IA doivent être encouragées, et la recherche de solutions d'IA dans des domaines d'intérêt mondial, comme par exemple les objectifs de développement durable, doit être favorisée. Dans l'idéal, les systèmes d'IA devraient être utilisés au profit de tous les êtres humains, y compris les générations futures.

Responsabilité

L'exigence de responsabilité complète les exigences ci-dessus et est étroitement liée au principe d'équité. Elle nécessite la mise en place de mécanismes pour garantir la responsabilité et l'obligation de rendre compte des systèmes d'IA et de leurs résultats, tant avant qu'après leur développement, leur déploiement et leur utilisation.

Annexe n°3 : Commission Européenne, *Livre blanc sur l'intelligence artificielle. Une approche axée sur l'excellence et la confiance*, 2020

Dans ce livre blanc, la Commission Européenne rend compte des axes prioritaires retenus pour l'encadrement à venir de l'intelligence artificielle (IA). Elle entend ainsi positionner l'Union Européenne comme un leader mondial du domaine de l'IA.

Pour cela, la Commission estime qu'un écosystème de confiance est nécessaire, ce qui implique un cadre réglementaire adapté à l'intelligence artificielle. Cette dernière est déjà soumise à la législation européenne sur les droits fondamentaux, à la protection des consommateurs et à la sécurité et la responsabilité des produits. Toutefois, certains composants de l'IA (comme l'opacité) peuvent rendre l'application de ces législations peu efficace. La Commission souligne en outre que si des bénéfices seront apportés par l'IA, celle-ci peut également nuire aussi bien matériellement que non-matériellement.

Sur les dangers liés aux droits fondamentaux, la Commission estime qu'ils peuvent résulter de défauts dans la conception globale des systèmes d'IA ou de la non-correction de biais. Les biais et la discrimination sont des risques inhérents à toute activité sociale ou économique et la prise de décision humaine n'est pas à l'abri de telles erreurs et biais. Cependant, le même biais, lorsqu'il est présent dans l'IA, pourrait avoir un effet beaucoup plus important, en touchant et discriminant de nombreuses personnes sans les mécanismes de contrôle social qui régissent le comportement humain. De telles erreurs et biais sont également susceptibles d'apparaître lorsque le système d'IA apprend, ce qui peut générer des erreurs dans les résultats de l'IA. De plus, l'opacité de l'IA augmente la difficulté de vérifier sa conformité avec les réglementations existantes. Les autorités chargées de l'application de la loi et les personnes concernées peuvent ne pas avoir les moyens de vérifier la façon dont une décision est prise et si les règles élémentaires ont été respectées.

Sur les risques pour la sécurité et le fonctionnement du régime de responsabilité, l'IA présente également des risques pour la sécurité des utilisateurs, lorsqu'elle est intégrée dans des produits et services. La Commission européenne prend l'exemple des voitures autonomes, où l'identification erronée d'un objet sur la route peut provoquer un accident engendrant des

blessures et des dommages matériels. Les risques liés à la sécurité, même s'ils ne se limitent pas à l'IA, peuvent être accrus avec cette technologie, et peuvent être causés par des défauts dans la conception de l'IA. Lorsque ces dangers se réalisent, la difficulté sera de prouver le caractère défectueux du produit basé sur l'IA à l'origine de l'accident. Pouvoir comprendre et pouvoir retracer les décisions prises par l'IA ayant mené à cet incident est alors nécessaire, notamment pour permettre aux éventuelles victimes d'obtenir réparation du préjudice subi.

Pour tenir compte de ces nouveaux potentiels risques, la Commission considère que la législation devrait évoluer, pour faire face aux situations suivantes :

- Assurer une application et une mise en œuvre efficaces de la législation communautaire et nationale existante ;
- Encadrer la limitation du champ d'application de la législation communautaire existante ;
- Tenir compte de l'évolution des fonctionnalités des systèmes d'IA ;
- Lever l'incertitude quant à la répartition des responsabilités entre les différents opérateurs économiques dans la chaîne de distribution ;
- Faire évoluer la notion de sécurité.

Afin d'y adosser un encadrement juridique pertinent, il importe donc en priorité de définir avec précision le champ d'application de cette stratégie, eu égard aux nombreux mythes planant sur le concept même d'intelligence artificielle et à l'absence de consensus scientifique sur la notion. Cependant, si la Commission insiste sur l'exigence d'une définition "*suffisamment souple pour tenir compte des progrès techniques tout en étant suffisamment précise pour garantir la sécurité juridique nécessaire*", elle n'entend pas faire naître un nouveau cadre qui créera une charge disproportionnée, en particulier pour les PME. Elle préconise ainsi que le nouveau cadre réglementaire ne s'applique qu'aux IA jugées à "haut-risque". Une qualification qui reposerait sur deux critères cumulatifs :

- Si l'IA est utilisée dans un secteur où, compte tenu des caractéristiques des activités généralement effectuées, l'on peut s'attendre à des risques importants, notamment du point de vue de la protection de la sécurité, des droits des consommateurs et des droits fondamentaux. Ce premier critère garantit que l'intervention réglementaire soit ciblée sur

les domaines où, d'une manière générale, les risques sont jugés les plus susceptibles de se produire. Les secteurs couverts devront être énumérés de manière spécifique et exhaustive dans le nouveau cadre réglementaire. Par exemple, les soins de santé, les transports, l'énergie et certaines parties du secteur public. La liste devrait être revue périodiquement et modifiée si nécessaire en fonction des évolutions pertinentes dans la pratique ;

- Et si l'utilisation de l'IA comporte des risques significatifs. L'évaluation du niveau de risque d'une utilisation donnée pourrait être basée sur l'impact sur les parties concernées. Par exemple, les utilisations d'IA qui produisent des effets pour les droits d'un individu ou d'une entreprise, qui présentent un risque de blessures, de décès ou de dommages matériels ou immatériels importants.

L'application de ces deux critères cumulatifs permettrait de garantir une sécurité juridique, car le champ d'application serait ciblé et proportionné. Mais la conception du futur cadre réglementaire pour l'IA nécessitera également de définir les obligations légales à imposer aux acteurs concernés par des IA à haut risque. C'est dans ce but que la Commission a établi une liste d'obligations que les opérateurs devraient suivre pour les IA à haut risque :

- S'assurer que les jeux de données utilisés pour les entraînements sont suffisamment exhaustifs et représentatifs pour envisager tous les scénarios, limiter la réalisation des risques, ou encore éviter les discriminations ;
- La tenue de registres sur les données utilisées pour entraîner les algorithmes, en y documentant les caractéristiques des jeux de données, les choix réalisés, les techniques et méthodologies de conception et d'entraînement ;
- Fournir certaines informations, notamment sur le recours à l'IA et son degré de fiabilité ;
- La robustesse et la précision, en prenant en compte dès la conception les risques de l'IA et en assurant une gestion des erreurs tout au long du cycle de vie de l'IA ;
- La supervision humaine, en faisant valider les résultats de l'IA par un être humain, en permettant une intervention en temps réel et en laissant la possibilité de désactiver l'IA.

La Commission européenne préconise que ces obligations soient supportées par l'acteur de la chaîne le plus à même de faire face au risque potentiel.

Annexe n°4 : Approche et conclusion du rapport de la CNIL, 2017[119]

- **Approche**

Le projet de loi sur la République numérique de 2016 a confié à la Commission nationale de l'informatique et des libertés (CNIL) la mission d'animer le débat sur les questions éthiques et sociétales soulevées par le développement rapide des technologies numériques.

En 2017, la CNIL a décidé d'axer ces discussions sur les algorithmes à l'ère de l'intelligence artificielle. À notre insu, ceux-ci s'insinuent de plus en plus dans tous les recoins de notre vie : résultats des moteurs de recherche sur le web, ordres financiers passés par les robots sur les marchés, diagnostics médicaux automatisés, répartition des étudiants candidats aux universités. Dans tous ces domaines, les algorithmes sont à l'œuvre. En 2016, le sujet des algorithmes s'est précipité de manière inédite sur la scène du débat public, suscitant une large couverture médiatique (questions sur l'algorithme du portail en ligne d'admission aux universités "Admission Post-Bac", l'utilisation de l'intelligence artificielle dans la stratégie de campagne électorale de Trump, le rôle des médias sociaux dans la diffusion de fausses nouvelles).

La réflexion éthique concerne les choix de société décisifs. Elle ne doit pas se faire sans tenir compte de cette dimension pluraliste et collective, surtout lorsqu'elle porte sur une question aussi transversale, qui touche à tous les aspects de notre vie sociale et personnelle. Il ne serait tout simplement pas possible de réunir au sein d'un seul comité toutes les compétences et perspectives nécessaires pour examiner les questions soulevées par les algorithmes dans des secteurs aussi variés que la santé, l'éducation, le marketing, la culture, la défense et la sécurité par exemple.

Aussi, plutôt que de mener directement des discussions centralisées sur ces sujets, la CNIL a décidé d'adopter une position originale en tant que leader d'un débat public national ouvert et décentralisé. Lors d'un lancement organisé le 23 janvier 2017, elle a ainsi appelé tous les acteurs et organisations intéressés (institutions publiques, société civile, entreprises) à organiser un débat ou un événement sur le sujet, dont ils feraient ensuite rapport à la CNIL. L'objectif était donc de

[119] Plus d'information sur ce lien : https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf

recueillir auprès des acteurs de terrain les sujets éthiques identifiés à ce jour ainsi que leurs idées pour les aborder.

Soixante partenaires se sont présentés en réponse à l'appel lancé par la CNIL, issus de secteurs très différents et représentant des structures diverses. Parmi eux, on peut citer la Ligue de l'Enseignement, la Fédération Française des Assurances (FFA), le Ministère de la Culture (DGMIC), Open Law (association de réflexion sur la justice) ainsi que des syndicats tels que la CFE-CFC et FO Cadres (pour le recrutement et les RH), etc.

Ils ont organisé 45 événements entre mars et octobre 2017 dans plusieurs villes françaises (ainsi qu'à l'étranger grâce à la Future Society de la Harvard Kennedy School), auxquels quelque 3 000 personnes ont participé. La CNIL a assuré une coordination globale et cohérente des événements.

Les manifestations organisées pour le débat public ont également été l'occasion de faire réfléchir la société française dans son ensemble sur des questions pour lesquelles la sensibilisation de tous nos contemporains, et pas seulement des experts, est de la plus haute importance démocratique et civique.

Conclusion de la CNIL

Les principes et recommandations énoncés à la fin de ce rapport sont le résultat de la synthèse que la CNIL a établi à partir des réflexions menées lors du débat public national qu'elle a animé de janvier à octobre 2017, avec l'aide de soixante partenaires.

Les recommandations politiques ont été exprimées de manière très large, en faisant appel au plus grand nombre possible d'acteurs publics et privés. Face aux défis posés par les algorithmes, l'ensemble de la société civile (citoyens, entreprises et associations) doit s'impliquer, être attentif et s'interroger dans un monde complexe. L'intention n'était donc pas de dire que la loi pouvait être le seul moyen approprié pour les appliquer. Bien au contraire, la plupart des recommandations peuvent être interprétées comme pouvant se concrétiser soit par un cadre

juridique contraignant, soit par une adoption volontaire de la part des parties prenantes des solutions englobant à des degrés divers ces deux extrêmes étant possibles.

Les discussions ont mis en évidence deux principes fondateurs, et comme certains des principaux défis éthiques posés par l'intelligence artificielle peuvent être regroupés sous ces principes, ils méritent une attention particulière.

Tout d'abord, le principe substantiel d'équité des algorithmes, qui s'appuie sur le principe déjà proposé par le Conseil d'État français (voir la section "Le principe d'équité"). Cette version intègre l'idée d'équité envers les utilisateurs, non seulement en tant que consommateurs mais aussi en tant que citoyens, et même envers les communautés dont le mode de vie pourrait être affecté par les algorithmes, que ceux-ci traitent ou non des données personnelles.

Deuxièmement, un principe plus méthodologique est celui de l'attention et de la vigilance continues. Il ne s'agit pas d'une vague incantation, mais d'une réponse étayée à trois défis centraux de la société numérique. Premièrement, la nature changeante et imprévisible des algorithmes à "l'ère du *machine learning*". Deuxièmement, la mentalité de cloisonnement qui affecte l'organisation des chaînes algorithmiques et qui conduit à une action menée de manière isolée, à l'indifférence aux impacts globaux du système algorithmique et à une diminution de la responsabilité. Troisièmement, le risque d'une confiance excessive dans les machines, qu'une forme de biais cognitif humain nous amène à considérer comme étant à l'épreuve des défaillances et sans biais. Le principe d'attention et de vigilance continues vise essentiellement à organiser l'état d'alerte permanent que nos sociétés doivent adopter à l'égard des objets socio-techniques complexes et changeants que représentent les chaînes ou systèmes algorithmiques. Cet état d'alerte implique de se soumettre constamment à un examen minutieux, à un doute méthodique. Cela concerne en premier lieu les individus, qui constituent les maillons des chaînes algorithmiques : il faut leur donner les moyens d'être à l'affût, de manière consciente et active, en cherchant toujours des réponses, dans cette société numérique. Mais cela concerne aussi les autres acteurs clés de notre société : les entreprises, bien sûr, pour modéliser des systèmes algorithmiques vertueux, ainsi que d'autres.

Grâce à l'approche universelle par laquelle ils sont nés, ces principes pourraient faire partie d'une nouvelle génération de principes et de droits de l'homme à l'ère numérique : une génération qui, après celles des droits-libertés, des droits de propriété et des droits sociaux, serait celle des "*system-rights*" organisant la dimension qui sous-tend notre monde numérique. N'ont-ils pas vocation à être défendus comme des principes généraux de la gouvernance mondiale de l'infrastructure de l'Internet ?

A l'heure où la France et l'Europe définissent leurs positions en matière d'intelligence artificielle, la question est tout à fait pertinente.

Annexe n°5 : Introduction à l'IA

L'apprentissage statistique est un ensemble de méthodes qui permet d'extraire l'information d'une base de données, en vue d'accomplir une tâche. Les tâches peuvent être très variées, comme par exemple la classification, la transcription, la traduction, la détection d'anomalie, le débruitage … La résolution de ces tâches s'articule autour de trois paradigmes : l'apprentissage *supervisé*[120] *ou non-supervisé*[121], la *représentation* du signal, et le choix du modèle *génératif* ou *discriminant*.

- **Apprentissage supervisé ou non supervisé**

L'apprentissage non supervisé correspond à la recherche de structure dans une base de données quelconque. C'est l'approche la plus générale. Disposant d'un grand nombre de *features (Propriétés, caractéristiques)* sur la cible de l'apprentissage et d'un grand nombre d'exemples, on en cherche les propriétés essentielles. On peut par exemple chercher à estimer la distribution de probabilité d'un signal, pour des tâches de débruitage ou de restauration. On trouve aussi dans cette catégorie le *clustering*, c'est-à-dire la recherche de groupes d'exemples similaires.

L'apprentissage supervisé suppose que chaque exemple soit associé à une catégorie cible. On guide l'apprentissage à partir d'exemples qui contiennent la catégorisation (on dit qu'ils sont *annotés*), afin d'obtenir une prédiction de la cible sur de nouveaux exemples. Ainsi, on peut entraîner un modèle supervisé à partir d'une base de données de photos de chats et de chiens, où chaque photo est associée à une des deux catégories. Historiquement la reconnaissance de chiffres manuscrits, entraînée à partir d'une grande quantité de photos de chiffres manuscrits, annotés de 0 à 9, a marqué une avancée de l'apprentissage supervisé, et les premiers résultats

[120] Tâche d'apprentissage automatique consistant à apprendre une fonction de prédiction à partir d'exemples annotés
[121] L'apprentissage non supervisé consiste à apprendre à un algorithme d'intelligence artificielle (IA) des informations qui ne sont ni classées, ni étiquetées, et à permettre à cet algorithme de réagir à ces informations sans supervision. Ce type d'apprentissage est principalement utilisé pour effectuer de la classification (*clustering*).

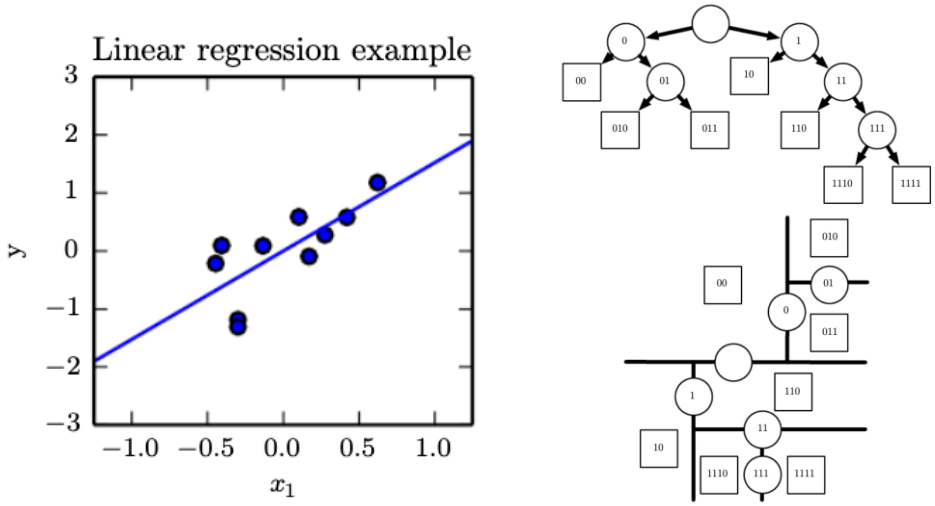
convaincants des modèles de *deep learning*. Le succès des approches supervisées repose donc sur l'existence de bases de données annotées, où chaque exemple contient la catégorie cible. La reconnaissance d'objets en traitement d'image se fait grâce à la mise à disposition de bases de données gigantesques : à titre d'exemple, ImageNet contient 3.2 millions d'images, annotées hiérarchiquement en 5247 catégories. La dépendance des algorithmes à ces données a créé de nouveaux métiers, qu'on appelle parfois "les micro tâches de l'IA", avec notamment le Mechanical Turk d'Amazon.

En pratique un système complet oscille souvent entre supervisé et non supervisé. On parle alors parfois d'apprentissage semi-supervisé, ou faiblement supervisé. L'acquisition de données annotées est parfois chère (certaines données bio-médicales passent par des prélèvements) et toujours limitée en nombre. La propagation d'annotations à des données similaires non annotées par exemple, est souvent une étape clé.

- **Régression, classification[122], arbre, neural network**

Nous présentons ici quelques exemples de tâches courantes en apprentissage.

[122] Attribuer une classe / catégorie à chacune des observations d'un jeu de données est une classification. Cette tâche se fait a posteriori, une fois les données récupérées.



Source : *Deep Learning*, Aaron Courville, Ian Goodfellow et Yoshua Bengio

La régression est la prédiction d'une valeur numérique en fonction d'une ou plusieurs entrées. Dans le cas d'une régression linéaire le problème consiste à trouver le vecteur de paramètres *w* qui permette de prédire la valeur *y* en fonction des entrées *x* en supposant une relation linéaire *y=w.x*. On définit la solution comme celle qui minimise l'erreur entre les observations et la prédiction. La régression linéaire est un exemple simple et très limité, mais il permet de comprendre comment un algorithme d'apprentissage fonctionne. On peut penser *w*. comme des poids qui déterminent comment chaque entrée contribue à la prédiction *y*. Si le poids *wi* est positif, une augmentation de la valeur de l'entrée *xi* implique une augmentation de la prédiction, et inversement. Si le poids d'une entrée *xi* est nul ou presque, alors l'entrée correspondante a une faible influence sur le résultat.

Un autre exemple de classification supervisée : les arbres de décision. Chaque nœud est associé à une région de l'espace d'entrée, et les feuilles de l'arbre décomposent l'espace en sous régions disjointes.

Sur un exemple précis : l'arbre est en haut de la figure, et l'espace des entrées en bas. Les nœuds de l'arbre sont en cercle et les feuilles carrées. Une valeur d'entrée passe par le premier nœud et descend jusqu'aux feuilles. Chaque nœud est un test binaire qui divise l'espace des entrées

en deux, si bien que l'espace est découpé en régions par l'ensemble de l'arbre, une région par feuille.

La rigidité de la division de l'espace obtenue limite les capacités de généralisation des arbres de décision. Chaque région requiert un ensemble de données d'apprentissage à part entière. La supériorité des réseaux de neurones vient de la capacité à apprendre d'un grand ensemble de régions avec beaucoup moins de points (Bengio and Monperrus, 2005). Par opposition aux arbres de décision on peut apprendre 2^k régions différentes avec seulement k données d'apprentissage.

Les réseaux de neurones profonds (*deep neural networks*) sont dédiés à l'estimation d'une fonction de transformation entre les entrées et les sorties. On parle de réseaux de neurones parce qu'ils sont constitués de plusieurs fonctions élémentaires appelées neurones, connectées en plusieurs couches. Le nombre de couches correspond à ce qu'on appelle la profondeur du réseau.

- **Détecter les structures : apprentissage de représentations**

La résolution des problèmes d'apprentissage dépend de la représentation choisie. On ne fournit pas à l'algorithme toutes les informations dont on dispose sur le patient, mais plutôt des statistiques très précises, comme la présence de cicatrices, de maladies, ou l'âge. Le choix des variables fournies en entrée a été déterminant dans le succès de la méthode. Au vu de la puissance de calcul de l'époque, le même algorithme aurait été incapable de fournir la même prédiction à partir de données plus volumineuses, comme des images ou des mesures de signaux continus. Le *feature design* correspond à l'étape de sélection ou de construction de descripteurs des signaux d'entrée, pour aiguiller l'apprentissage vers le résultat.

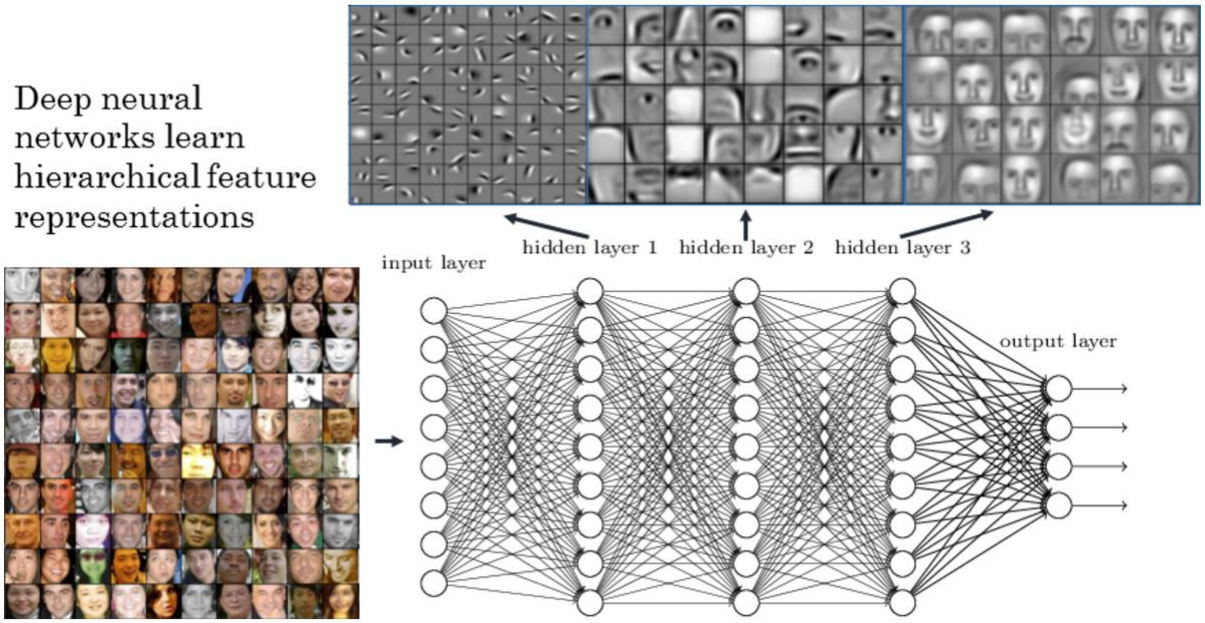
Les descripteurs d'image sont nombreux : les histogrammes de couleurs ont été utilisés pour des détecteurs de contexte (photo/non photo, extérieur/intérieur, nudité). Les gradients orientés, en calculant les différences des pixels avec leur voisin, servent d'approximation des

contours des objets. Associés à des classifieurs peu profonds, ils peuvent fournir des résultats satisfaisants. Une première étape de construction de *feature* peut chercher à obtenir des informations comme la direction de la lumière, la segmentation *foreground/background*, qui sont ensuite agrégés dans des vecteurs de taille fixe (par exemple les *visual bag of words*, ou les vecteurs de Fisher).[123][124]

En traitement de texte, l'ensemble des fonctions de comptage (TF IDF, concurrences, ...) sont suffisantes pour la classification des mails en spam/non-spam. En traitement du signal sonore la construction de variables liées à la perception humaine (hauteur de note, cochléogramme, réverbération, statistiques globales...), sont encore à la base d'algorithmes performants de reconnaissance de locuteur par exemple. Ces caractéristiques sont parfois choisies parce qu'elles sont faciles à calculer (*zero crossing rate* en audio, les gradients en image, ...) et ne correspondent pas à des séparateurs optimaux pour les tâches de classification. Elles correspondent aussi selon les cas à des connaissances de métier, propres à la nature des données et à la tâche à accomplir. En amont des performances statistiques, certaines résistent encore au *feature learning*, mais dans des cas qui font de plus en plus figure d'exception. L'apprentissage de représentation remplace l'étape de *feature design* dans la plupart des approches modernes.

[123] J. Sivic, & A. Zisserman, *Video Google: A text retrieval approach to object matching in videos, 2003*. In *null* (p. 1470). IEEE.

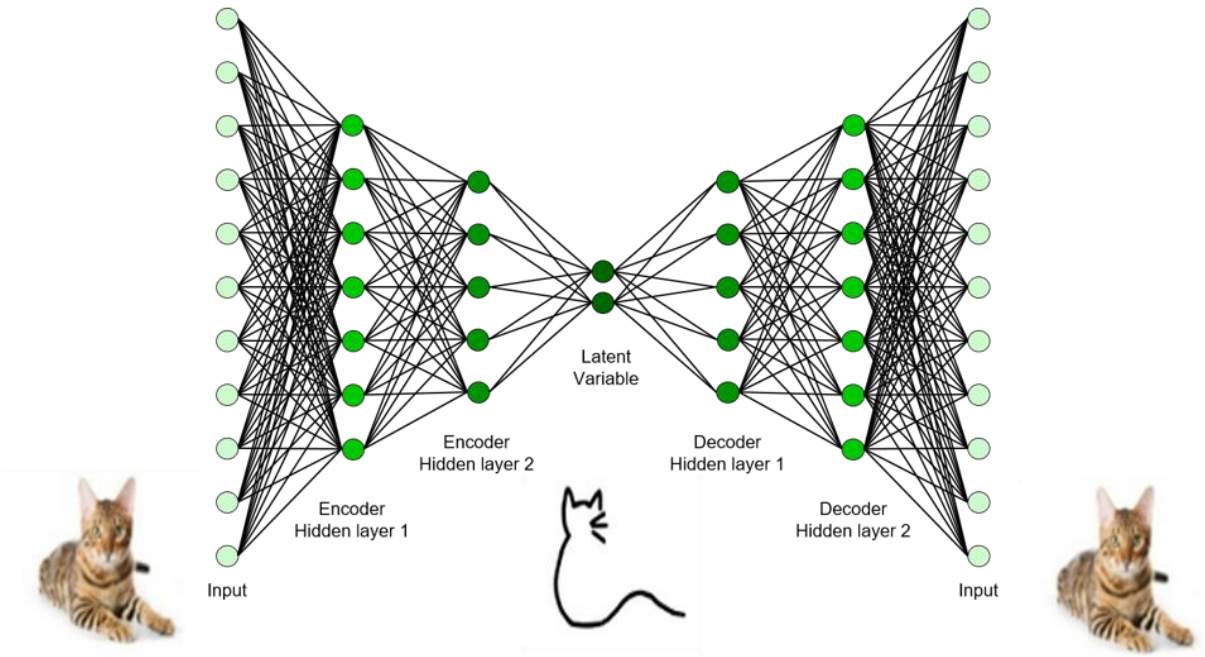
[124] F. Perronnin, J. Sánchez, & T. Mensink, *Improving the fisher kernel for large-scale image classification. In European conference on computer vision* (pp. 143-156), 2010. Springer, Berlin, Heidelberg.



Source : LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* 2015, 521, 436–444.

Pour beaucoup de tâches, le choix des représentations est resté un problème difficile, notamment en image. La reconnaissance d'objet sur des photos basées sur une description géométrique minutieuse n'a jamais rien donné. Il s'agit de trouver une représentation robuste à toutes les variations d'angle, de décor que comporte la classe des photos d'un objet donné. L'avènement des réseaux de neurones vient de leur capacité à justement apprendre la représentation des données. Et ce n'est souvent qu'une fois l'apprentissage terminé qu'on arrive à identifier la représentation qu'a choisi le réseau. Au lieu de décrire le signal, l'image, on laisse un réseau de neurones apprendre en même temps que la prédiction de la cible, la représentation du signal brut. D'ailleurs on reconnaît parfois dans les couches de certains réseaux de *deep learning* des *features*. Dans la première couche des réseaux convolutifs pour le traitement d'une image, on reconnaît des filtres de gradient (reconnaissance des gradients horizontaux par exemple), ce qui est comparable au traitement des cellules V1 du cerveau humain (les premières à traiter l'info visuelle). Dans le cas de Wavenet, un réseau dédié au traitement du signal audio, celui-ci apprend directement sur les formes d'ondes plutôt que sur

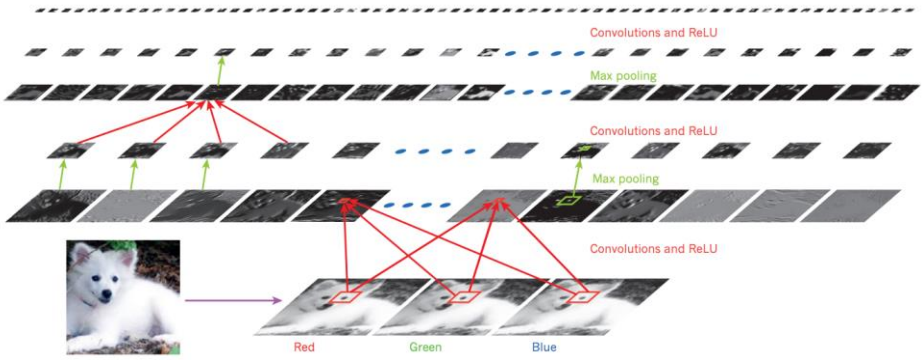
des spectrogrammes. Après convergence, on retrouve dans les premières couches la construction du plan temps fréquence. Il s'en suit une plus grande complexité en termes de nombre de paramètres, et donc d'une augmentation des besoins en puissance de calcul.



Source : *Autoencoder* By Prof. Seungchul Lee iSystems Design Lab

Une façon de comprendre l'apprentissage de *features* est l'étude des auto-encodeurs. Un auto-encodeur est un réseau de neurones qui tente de générer en sortie le signal d'entrée. Entre les deux, on place une couche de paramètres, de dimension inférieure au signal d'entrée. Ainsi l'algorithme est contraint de trouver les critères minimaux pour reconstruire le signal d'entrée à partir d'un plus petit nombre de variables : l'idée est que la structure la plus pertinente de la donnée est concentrée autour d'un espace (linéaire ou non) de dimension plus petite. Dans le cas d'une seule couche de neurones à activation linéaire, on retrouve une réduction de dimension classique, les poids des nœuds convergent vers une PCA. L'intérêt vient de l'information qu'on peut ajouter via la structure des réseaux intermédiaires, et ainsi d'apprendre des espaces latents sur des variétés non linéaires.

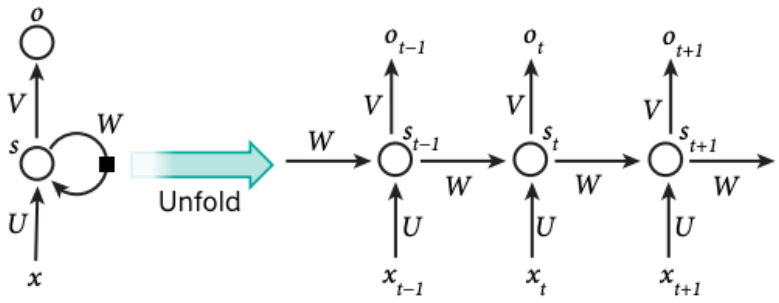
- **Les réseaux convolutifs et récurrents**



Source : LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* 2015, 521, 436–444.

Les réseaux de neurones convolutifs sont construits pour traiter des données structurées en matrices. Par exemple, une image en couleur est représentée par trois tableaux 2D. L'organisation du réseau est basée sur la notion de connectivité locale entre un grand nombre de couches. Chaque couche est connectée localement aux paramètres des couches précédentes, ce qui permet une organisation hiérarchique de l'information : les couches les plus proches des données d'entrées réagissent directement aux pixels des images, pour détecter des contours et des motifs, les motifs s'agrègent pour former des objets, et les couches les plus hautes détectent des aspects globaux. Cette structure hiérarchique existe en traitement de texte, les phonèmes induisant des syllabes, des mots, puis des phrases.

En pratique chaque couche est composée typiquement de nœuds convolutifs, une fonction d'activation non linéaire, et d'un nœud de *pooling* ou de mise en commun. Le *pooling* réduit la complexité, et représente l'invariance du système aux changements locaux.



Source : LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* 2015, 521, 436–444.

Les réseaux récurrents sont construits pour prendre en compte l'aspect séquentiel des données, notamment en traitement de la parole et du texte. Les *Recurrent Neural Networks* (RNN) fonctionnent avec une mémoire : ils traitent les données d'entrée une par une en retenant dans un des nœuds l'information des données précédentes. Ils sont particulièrement performants pour les tâches d'autocomplétion de mots ou de phrases.

- **Modèle discriminant vs modèle génératif**

Les modèles discriminants apprennent à décrire la frontière entre les catégories cibles, alors que les modèles génératifs modélisent la distribution des exemples de chaque catégorie. Par exemple, si on connaît la nature Gaussienne d'un signal, un algorithme qui apprend les paramètres de cette distribution peut être utilisé pour générer de nouveaux exemples qui paraissent plausibles au regard des données d'entrées. L'idéal pour un algorithme génératif est justement de générer des exemples impossibles à distinguer des données d'entrées.

Annexe n°6 : Fonctionnement des algorithmes de création de contenu

- **Les réseaux antagonistes génératifs (GANs)**

Les *GANs* se différencient des autres types de réseaux tout d'abord par leur but qui est la création de contenus nouveaux, que ce soient des images, du texte de facture artistique ou scientifique, de la vidéo, des peintures ressemblant aux maîtres, etc. En soi, ce but est plus difficile à quantifier que, par exemple, les applications à la classification supervisée décrites aux chapitres précédents. Néanmoins des techniques ont été mises en place avec de bons résultats. Les briques de base sont toujours les réseaux profonds mais l'architecture est différente.

Nous allons nous attarder plus en détail sur une catégorie de réseaux dits *GAN* (pour *Generative Adversarial Network*, qui pourrait se traduire par "Réseaux Génératifs Antagonistes") introduits en 2014 par Ian Goodfellow et ses collaborateurs.

L'idée est de faire interagir deux réseaux : un réseau dit "générateur" (G) qui s'entraîne à proposer des nouvelles images et un autre, le "discriminateur" (D) qui dispose d'une base préexistante d'images réelles et qui s'entraîne à distinguer les images réelles de celles générées par (G). La suite d'interactions est la suivante : d'abord (G) propose quelques images et (D) les reçoit avec leurs étiquettes ("faux") et s'entraîne à les distinguer des images choisies au hasard dans la base réelle (étiquette "vrai"). Pendant la deuxième phase de l'entraînement c'est le réseau (D) qui est figé et le réseau (G) qui est entraîné à créer des images qui, une fois passées par l'analyse du réseau (D) (actuellement figé) donnerait la réponse "image réelle" (alors qu'il s'agit de créations de (G)).

Ensuite le processus est répété : (D) prend à nouveau des images, maintenant un peu meilleures, de (G) avec l'étiquette "faux" et des images réelles avec l'étiquette "vrai" et s'entraîne à les distinguer d'une manière automatique ; ensuite c'est à nouveau le tour de (G) et ainsi de suite. La figure ci-dessous donne une illustration de ce processus.

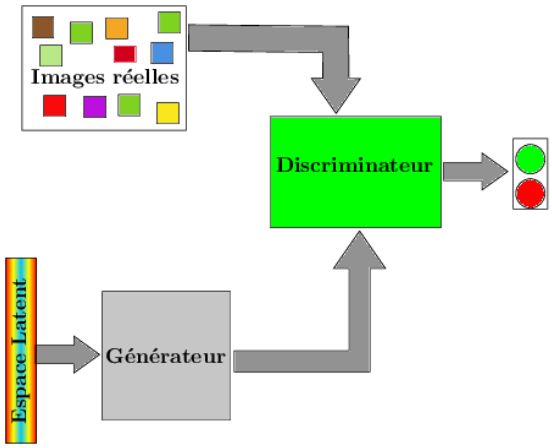
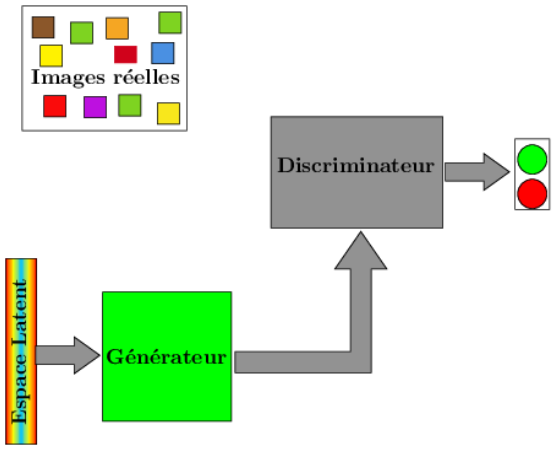
Phase d'apprentissage pour (D)	Phase d'apprentissage pour (G)
	

Illustration du fonctionnement d'un *GAN*. En gris les parties qui ne sont pas actives, en vert les parties actives. Le *GAN* alterne entre deux phases : à gauche une phase dans laquelle le discriminateur est actif et le générateur inchangé. Le discriminateur apprend à faire la différence entre image réelle et générée. Lors de la phase suivante (à droite) c'est le tour du générateur de s'entraîner avec un discriminateur inchangé.

La description plus précise de la mécanique du processus de génération d'images nécessite d'introduire la notion d'espace latent ; de la même manière que le réseau (D) prend une image pour en rendre un verdict résumé dans une réponse OUI / NON, le travail de génération a en quelque sorte une tâche inverse : il prend en entrée un vecteur de nombres (dans les faits : pris au hasard selon une distribution connue, par exemple la loi normale) et les transforme en une image de la taille attendue. Chaque image sera donc déterminée par la donnée d'une petite quantité de nombres qui forment ce qu'on appelle l'espace latent. Chaque image générée a une description unique en termes des variables de l'espace latent. La dimension de cet espace est, généralement, plus petite que la dimension de l'image résultat afin :

- D’une part d'éviter de travailler avec beaucoup d'informations (tous les pixels d'une image ne sont pas indispensables ni complètement indépendants les uns des autres),
- D’autre part d’avoir tout de même une dimension pas trop faible afin de permettre la génération d'images assez diverses. L'utilisateur ne voit pas cet espace et d'ailleurs peu d'intuitions existent pour comprendre le rôle précis de chaque élément de l'espace latent dans la création de l'image finale.

Cette interaction entre un générateur et un discriminateur caractérise les *GANs*. La possibilité de tirer aléatoirement une image ou un son n’est, au premier regard, pas surprenante pour l’ingénieur puisqu’il s’agit de l’essence du numérique. En effet, dans un ordinateur classique toute donnée est représentée sous forme d’une suite de nombres binaires. La nouveauté est que nous sommes désormais capables de “tirer au hasard”, éventuellement conditionnellement dans cet espace latent.

La manière de procéder des *GANs* est en revanche plus étonnante pour le mathématicien : la distribution latente choisie est fréquemment une gaussienne unimodale tronquée et les réseaux transformant cette gaussienne en image sont continus. L’image d’un connexe par une fonction continue étant connexe, cela signifie que le support de la distribution des images générées par un *GAN* le sera aussi. Cette hypothèse semble trop forte pour des images naturelles. Il existe donc une forme de garantie qu’une partie des images générées seront de mauvaise qualité visuelle, même si en pratique des astuces sont mises en place pour aider le réseau à séparer l’espace d’arrivée en plusieurs blocs. Une technique fréquente est par exemple d’apprendre les *GANs* de manière conditionnelle à la classe de sujet à générer, ce qui contribue à la génération d’arrière-plans relativement peu complexes (flous, unis, motif végétal...).

- **Des GANs très divers**

Le protocole d'apprentissage des *GANs* connaît de nombreuses variations. Des propositions ont été faites afin d'améliorer la convergence, la qualité des résultats, la vitesse de convergence, etc. Ceci a eu pour conséquence d'établir toute une zoologie des réseaux de type *GAN*.

Malgré des résultats impressionnants, ces réseaux restent difficiles à apprendre : les descentes de gradient stochastique sur un *GAN* ont la réputation de ne jamais complètement converger et de ne pas toujours trouver un minimum satisfaisant. Par exemple, BigGAN dont l'entraînement est évalué à $50 000 en prix *cloud* public ne produit des modèles "*amenable to truncation*" (i.e. pour lesquels on sait *sampler* avec la qualité promise dans l'article initial) que 60% du temps.

En pratique, quand l'entraînement se passe bien, on observe une oscillation en partie générative et discriminative. Si à l'inverse l'une des deux devient plus rapidement performante que l'autre, l'apprentissage se retrouve généralement bloqué dans un minimum local. Schématiquement, si le discriminateur est devenu bon pour identifier les images de vrais "chats" alors le générateur à deux solutions : soit améliorer la production de ses chats pour qu'ils soient plus difficiles à identifier, soit ne plus jamais générer d'images de chats. Cette seconde situation connue sous le nom de "*mode collapse*" arrive fréquemment quand le discriminateur est vraiment trop performant : ses gradients deviennent non informatifs car variant trop brutalement.

Il est devenu fréquent de ne pas démarrer complètement au hasard. Les approches modernes entraînent le discriminateur et le générateur sur des tâches différentes. Par exemple, générer du texte nécessite probablement le même type de compétences que compléter des trous dans une phrase. C'est exactement l'approche empruntée par un modèle publié par Google sous le nom de *Bidirectional Encoder Representations from Transformers* (BERT)[125]. Le générateur y est pré-entraîné en masquant 15% des mots de la phrase initiale et en visant simplement à apprendre un RNN (Bidirectionnel) capable de deviner les mots manquants. Cette tâche, avec un masquage un tout

[125] Modèle de langage développé par Google en 2018. Cette méthode a permis d'améliorer significativement les algorithmes de traitement automatique de la langue.

petit peu plus sophistiqué qu'une simple sélection au hasard, est connue sous le nom de *Masked Language Model* (MLM).

Le pré-entraînement du discriminateur fait lui appel à une tâche connue sous le nom de "*Next Sequence Prediction*". L'idée est, à partir d'un ensemble de textes, de construire des couples de phrases et d'entraîner un modèle à déterminer si les deux phrases étaient ou non consécutives (en faisant tout de même attention à fournir des phases consécutives sur 50% des exemples). Ici encore il devient possible de tirer un avantage des techniques supervisées. De manière plus générale, l'apprentissage de tâches dont les étiquettes peuvent être obtenues sans intervention humaine est connue sous le nom "*self supervised learning*" et est considérée comme une des pistes les plus prometteuses pour l'année 2020 par de nombreux acteurs du domaine dont Y. Bengio.

La structure du réseau utilisé étant également variable, les textes ont tendance à se baser sur des réseaux récurrents, tandis que les images utilisent plus fréquemment des réseaux à convolution comme Radford, Metz, et Chintala en 2016 avec DCGAN (*Deep Convolutional Generative Adversarial Network*). Aujourd'hui des transformateurs sont de plus en plus fréquents pour les textes, tandis que des *U-net* se répandent dans les images, y compris en 3D. Dans tous les cas, il est fréquent d'intégrer au réseau des mécanismes d'attention. Ces mécanismes, qui trouvent leur origine dans l'observation de l'amélioration des performances en traduction Français/Anglais par l'inversion du sens de lecture à l'entrée du réseau, ont progressivement été améliorés. Aujourd'hui les réseaux construisent des représentations en se basant sur des mots situés indifféremment avant ou après, voire en ignorant automatiquement des mots. De manière encore plus surprenante, il semble que ces architectures soient désormais capables de redécouvrir les règles basiques de la grammaire et d'identifier des hiérarchies de concepts.[126]

Il existe d'autres contributions notables. Une version très appréciée et qui a permis d'améliorer la convergence a été l'introduction des "*Wasserstein GANs*" (abrégé WGAN) par Arjovsky, Chintala, et Bottou en 2017 qui proposent d'adapter un outil technique particulier qui est la distance entre lois de probabilités. Les *GANs* peuvent être vus comme la génération des instances d'une variable

[126] Voir : https://pair-code.github.io/interpretability/bert-tree/

aléatoire suivant une loi de probabilité inconnue mais donnée par quelques observations (exemples). Or, si on a plusieurs propositions de lois des probabilités, il faut être capable de distinguer quelle loi se rapproche le plus de celle dont viennent les observations. L'objet mathématique qui quantifie cette distinction est dit "distance statistique" ou distance entre lois de probabilités. Arjovsky, Chintala, et Bottou proposent d'employer une distance particulière, dite distance de Wasserstein. Ceci change légèrement le discriminateur (et son entraînement) en lui imposant des contraintes techniques mais a le potentiel d'améliorer la diversité d'images nouvellement générées. Cette proposition a été perfectionnée quelques mois après par la contribution de Gulrajani et al. (2017) qui ont introduit les "Wasserstein *GANs* avec pénalisation du gradient" (en abrégé WGAN-GP) où une certaine souplesse est laissée au discriminateur (les contraintes sont imposées d'une manière plus souple).

Une architecture très singulière (deux générateurs et deux discriminateurs) a été testée par Liu et Tuzel en 2016 avec de bons résultats, donnant naissance au coGAN (Coupled GANSs).

Un peu après, Karras et al. de NVIDIA ont constaté que si le discriminateur est trop bon le générateur se "décourage" et ne propose que des solutions de faible qualité. Afin de corriger ceci, ils ont inventé le paradigme du "*Progressive Growing Generative Adversarial Network*" (aussi appelé "Progressive *GAN*" ou encore "ProGAN") où le discriminateur et le générateur agissent au début seulement sur des versions de très faible résolution (par exemple 4x4 pixels). Ceci permet au générateur de pouvoir proposer des images synthétiques que le discriminateur aura du mal à distinguer des images réelles. Une fois que le générateur est bon à ce niveau de résolution, la qualité d'image est augmentée (8x8, et ainsi de suite) et une nouvelle étape commence. Ceci se passe plusieurs fois jusqu'à ce que le générateur propose des images de haute résolution (1024x1024 ou même plus) et de bonne qualité. C'est un des algorithmes les plus efficaces à ce jour. Toujours venant de NVIDIA, un autre *GAN*, le "*Style-Based Generative Adversarial Network*" (StyleGAN, Karras, Laine, et Aila 2019) s'attaque à un autre versant de la génération d'images : comment générer des images ayant des caractéristiques données (pose, couleur de cheveux, forme

du visage, couleur des yeux, caractéristiques de la peau, etc.) en associant à chaque trait un niveau de résolution différent.

Ceci ne clôt pas la liste des versions de *GANs* : il faut y inclure encore les "*Cycle-Consistent Generative Adversarial Network*" (CycleGAN, par Zhu et al. 2017), les "*Big Generative Adversarial Network*" (BigGAN, par Brock, Donahue, et Simonyan 2019), les Pix2Pix (Isola et al. 2018), etc. Et la recherche sur les *GANs* et leurs variantes continue d'avancer.

- **Les auto-encodeurs variationnels[127] *(Variational Auto Encoders - VAE)***

La qualité de la génération d'un *GAN* repose sur un équilibre délicat entre le générateur et le discriminateur. Si on fait une analogie avec le monde des peintres et des critiques d'art, on peut voir que la qualité des œuvres sera meilleure quand les peintres amateurs auront la chance d'améliorer leurs créations sans forcément être sévèrement réprimandés pour leurs erreurs de jeunesse. La critique doit aussi fournir des pistes d'amélioration afin que les idées nouvelles puissent se mettre en place. Or, dans les *GANs*, il arrive parfois que la critique (le "discriminateur") devienne trop forte, bloquant ainsi tout progrès. L'inverse peut aussi arriver, quand un peintre star monopolise toute l'attention. Dans les deux cas, le progrès est limité et dans les *GANs*, ceci conduit à des cas où la qualité des résultats n'est pas satisfaisante. Pour pallier ce défaut qui est inhérent à la structure des *GANs*, plusieurs chercheurs ont proposé d'autres architectures capables de générer du contenu nouveau. Nous allons en invoquer une, les "*Variational Auto Encoders*" (VAE). Un VAE est un *GAN* renversé : il commence par un étage dit d'encodage qui transforme l'entrée dans une (courte) liste de caractéristiques, dit "espace latent." À la suite de cela, un étage "décodeur" prend en entrée un exemple de caractéristiques et reconstruit l'exemple de départ. Il contient, comme les *GANs*, des ingrédients techniques tels que les "distances statistiques."[128]

[127] Un auto-encodeur, ou auto-associateur est un réseau de neurones artificiels utilisé pour l'apprentissage non supervisé de caractéristiques discriminantes. L'objectif d'un auto-encodeur est d'apprendre une représentation (encodage) d'un ensemble de données, généralement dans le but de réduire la dimension de cet ensemble. Un VAE est une variante générative et probabiliste de l'architecture de réseau de neurones auto-encodeur.
[128] Voir les travaux de Gabriel Turinici sur les distances statistiques https://arxiv.org/abs/1911.13135

Jusqu'ici un tel réseau ne fait au mieux qu'une compression d'informations (de la même manière que la procédure qui consiste à prendre une image numérique de plusieurs mégapixels et la transformer en une photo format JPEG de quelques kilo-octets). Le VAE va plus loin en proposant que l'ensemble des caractéristiques (l'espace latent) ait une structure standardisée et connue. Si jamais ceci est réalisé, alors à la place de partir d'une image réelle, on peut très bien insérer au milieu du VAE, entre l'encodeur et le décodeur, un élément nouveau de l'espace latent (on sait en générer car on a dit que celui-ci a maintenant une structure bien définie, ou en langage mathématique, nous connaissons la loi de probabilité sur celui-ci ; on peut imaginer l'espace latent comme l'intervalle [0,1] avec une loi de probabilité uniforme, il suffit alors de tirer au hasard un nombre entre 0 et 1; en pratique ce sera en fait une loi gaussienne multidimensionnelle). L'élément latent inséré pourra passer par le décodeur comme tout autre élément de la base initiale et donner ainsi naissance à du contenu nouveau, similaire aux données utilisées en entrée de l'encodeur et qui ne servent plus désormais dans l'étape de génération (les données ont servi à trouver le réseau optimal qui encode-decode fidèlement la base initiale ET a une structure latente prescrite).

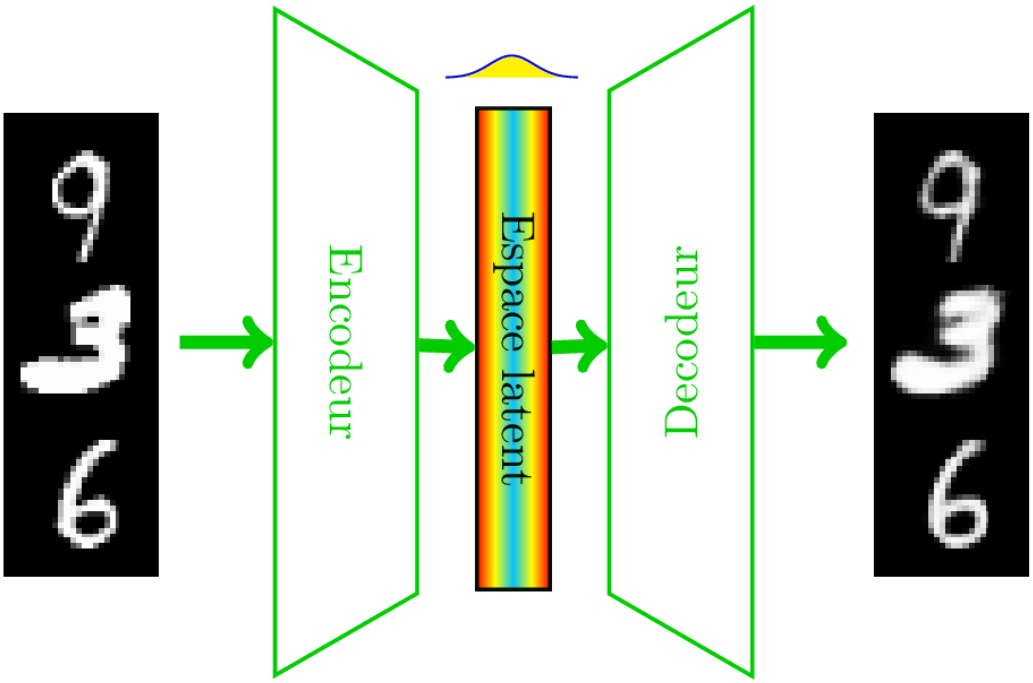
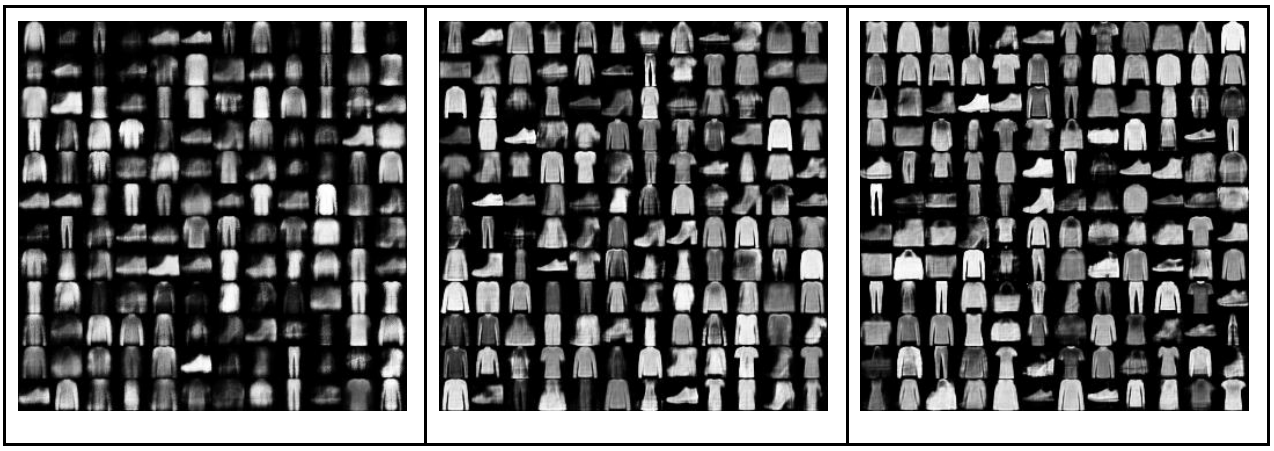


Figure : VAE : illustration de la structure d'un encodeur variationnel. Les images en entrée sont reconstruites en sortie, après passage par un "goulot d'étranglement" au milieu. Les réseaux Encodeur et Décodeur sont optimisés pour améliorer à la fois la qualité de la reproduction in fine mais aussi la concordance avec la distribution fixée à l'avance dans l'espace latent. Crédits : G.T.'19.

Convergence du VAE pour une base de données contenant des images de vêtements (dite FMNIST). A gauche : images générées après une seule lecture de la base (une "epoch"), au milieu après 20 passages et à droite après 500. On remarque une amélioration de la vraisemblance des images et de leur variété (évitant le danger de concentration spécifique au *GANs*) .



- **Modèles de flux**

Remplacer l'espace latent constitué d'une distribution des VAE par un simple espace Euclidien permet de redécouvrir une structure plus ancienne connue sous le nom d'auto-encodeur. Ces auto-encodeurs sont plus faciles à entraîner que les VAE (il s'agit d'une tâche de compression de contenu), mais il est difficile de les utiliser comme une distribution. L'espoir des modèles de flow est de tirer le meilleur des deux mondes.

L'apprentissage consiste à apprendre une transformation inversible depuis la distribution des données vers une distribution connue comme une loi normale. L'apprentissage peut donc se faire sans avoir à calculer une dérivée dans l'espace latent constitué de paramètres de lois. Générer une image se fait en tirant selon la loi connue et en utilisant la transformation inverse. Cette idée a été dérivée de différentes manières pour les images conduisant à Glow et FFJORD qui offrent des performances plus stables que les *GANs* mais restent un peu en retrait une fois l'optimisation des hyper-paramètres achevée. C'est en revanche une direction très active de recherche.

En collaboration avec :

Ekimetrics.