Les technologies du faux : un état des lieux

Par Christophe Deschamps

« Ceci est la réalité.
Et j'en fais partie. »
Philip K. Dick

Publication initiale le 15/11/2021 sur : **observatoire-strategique-information.fr**/2021/11/15/les-technologies-du-faux-un-etat-des-lieux

Introduction

C'est une évidence de rappeler que la manipulation, la tromperie, l'intoxication et plus globalement toutes les méthodes visant à influencer une cible sont aussi vieilles que l'humanité. Rhétorique, stratagèmes, artifices, publicité, *nudge marketing*,… si chaque époque et chaque culture ont développé leurs propres méthodes (tout en continuant d'exploiter les précédentes), l'objectif reste inchangé : modifier la perception qu'une "cible" a de la réalité afin de la faire agir en conséquence. Notre époque n'est pas avare en la matière et l'intelligence artificielle, dans ses composantes d'apprentissage machine (*machine learning*) et plus spécifiquement d'apprentissage profond (*deep learning*), lui confère une niveau d'incidence potentiellement considérable.

En liminaire, notons que la manipulation n'a pas nécessairement besoin de s'appuyer sur le faux pour être efficace et que s'appuyer sur le vrai, le véridique, la rend à la fois plus crédible et plus complexe à déconstruire. Là où le *fake* peut être mis à mal par une seule preuve pointant la volonté de tromperie et discréditant de fait son émetteur connu ou supposé, la manipulation, qui s'appuie sur une demi-vérité, est plus subtile et partant plus difficile à prouver. Pour autant le régime du faux, entraînant à sa suite celui du *fake*, étend tous les jours son emprise comme nous allons le constater. Précisons en effet avant d'aller plus loin que si le faux désigne ce qui n'est pas véridique, il ne porte pas d'intentionnalité et se différencie ainsi du *fake*, c'est à dire de ce qui est fabriqué dans l'objectif manifeste de tromper. S'appuyer sur des *fakes* est donc une modalité de l'influence, tout comme s'appuyer sur des demi-vérités (ou sur des arguments factuels et bien ordonnés)[1].

Évoquant ce thème en 2016, le chercheur François-Bernard Huyghe, écrivait à propos de l'émergence du web "2.0" dès 2005 : *"Le facteur qui va tout bouleverser est l'équation «numérique plus réseaux». (…) Si chacun peut devenir émetteur à son tour et non simple récepteur des mass médias(…) il peut informer donc désinformer."* Il évoquait corollairement la nécessité pour les réseaux sociaux de capter l'attention des cibles, ce qui passe par la possibilité de pouvoir modifier les contenus numériques "*à très faible coût, avec des exigences de plus en plus faible en termes de compétence techniques (logiciels plus simples et accessibles)*" et précisait enfin que "*les ressources documentaires, banques d'images, bases d'information en ligne, immédiatement, gratuitement… permettent de piocher dans des réserves de données qui permettent de forger des trucages vraisemblables. Le travail du faussaire est donc facilité pour ne pas dire banalisé.*"

Un discours qui anticipait bien la période actuelle puisqu'un an plus tard émergeait le phénomène des *deep fakes* (en français "hypertrucages"), capable de produire des vidéos ou des photos dans lesquelles, par exemple, le visage d'un homme politique est remplacé par un autre ou encore lui faisant remuer les lèvres, telle une marionnette, afin de lui faire dire ce que l'on souhaite. L'expression "*deep fake*" a été forgée en ajoutant au terme "*fake*" le "*deep*" de "*deep learning*" qui désigne un ensemble de techniques permettant à l'intelligence artificielle d'apprendre à reconnaître et reproduire des formes, structures, objets, visages et qui pour cela tire parti de bases d'images existantes. Le faussaire devient ainsi un faussaire augmenté par l'IA.

Historiquement, deux innovations ont permis l'émergence des *deepfakes*, le système de reconnaissance facial développé par Yann LeCun pour Facebook, baptisé

Deepface[2], et les GAN (*Generative Adversarial Network*) développés par le chercheur Ian Goodfellow[3] dans ses recherches en intelligence artificielle pour le compte d'Alphabet (Google)[4]. Les *deepfakes* sont générés par les GAN , une technique dans laquelle deux algorithmes, le générateur et le discriminateur, sont en compétition, le premier produisant des faux de plus en plus crédibles a mesure que le second les détecte. Le "dialogue" étant entretenu jusqu'à obtenir des faux plausibles pour l'œil humain. Par défaut, un GAN est capable de produire une image aléatoire à partir de n'importe quel jeu d'images avec lequel on l'alimente. Afin qu'il soit en mesure de recréer l'image d'une personne spécifique il faut donc introduire une condition, on parle alors de "*conditional GAN*" ou cGAN (voir par exemple l'algorithme *pix2pix*), c'est à dire l'entraîner sur un set d'images de la personne que l'on tente de récréer, ce qui implique a priori que cette personne soit suffisamment populaire ou visible pour que l'on puisse collecter des photos d'elle.

Des visages et des corps

Sans surprise, les premiers *deepfakes* apparus à l'automne 2017 étaient des vidéos pornographiques dans lesquelles les visages d'actrices connues (Gal Gadot, Emma Watson, Jennifer Lawrence,…) remplaçaient ceux des actrices originales[5]. Les outils *open source* utilisés s'inspiraient alors de *face2face*, un logiciel expérimental présenté par une équipe de chercheurs en 2016[6].

En quelques années ces technologies se sont commoditisées et leurs usages démultipliés. Outre les changements de visages sur des vidéos ou des photos, qui commencent à être utilisés dans le cinéma (cf. l'apparition de Luke Skywalker jeune *dans un épisode de The Mandalorian*), on a vu arriver de nombreuses autres possibilités proposées par des développeurs indépendants, des équipes universitaires mais aussi, de plus en plus souvent, des entreprises.

Hormis les vidéos truquées d'hommes politiques ou d'acteurs, les *deepfakes* ont commencé à entrer dans les usages via des applications pour smartphone utilisables par tous. C'est le cas par exemple de *Reface*, de *Cupace* ou de *Hellos*, des applications de "*face swapping*" pour smartphone qui permettent d'incruster un visage dans une scène de film, un tableau, un clip vidéo,… D'autres applications familiarisent également l'utilisateur avec les GAN comme *FaceApp* pour aider au *relooking, Photo Glory* ou *Deoldify* qui offrent la possibilité de coloriser des photos noir et blanc, *Unfade* qui permet de les restaurer, ou encore *Deep Nostalgia* qui permet d'en animer les visages.

Par ailleurs des services permettant de générer de faux visages particulièrement crédibles sont en ligne depuis plusieurs années déjà comme par exemple *Thispersondoesnotexist.com* ou *generated.photos*.

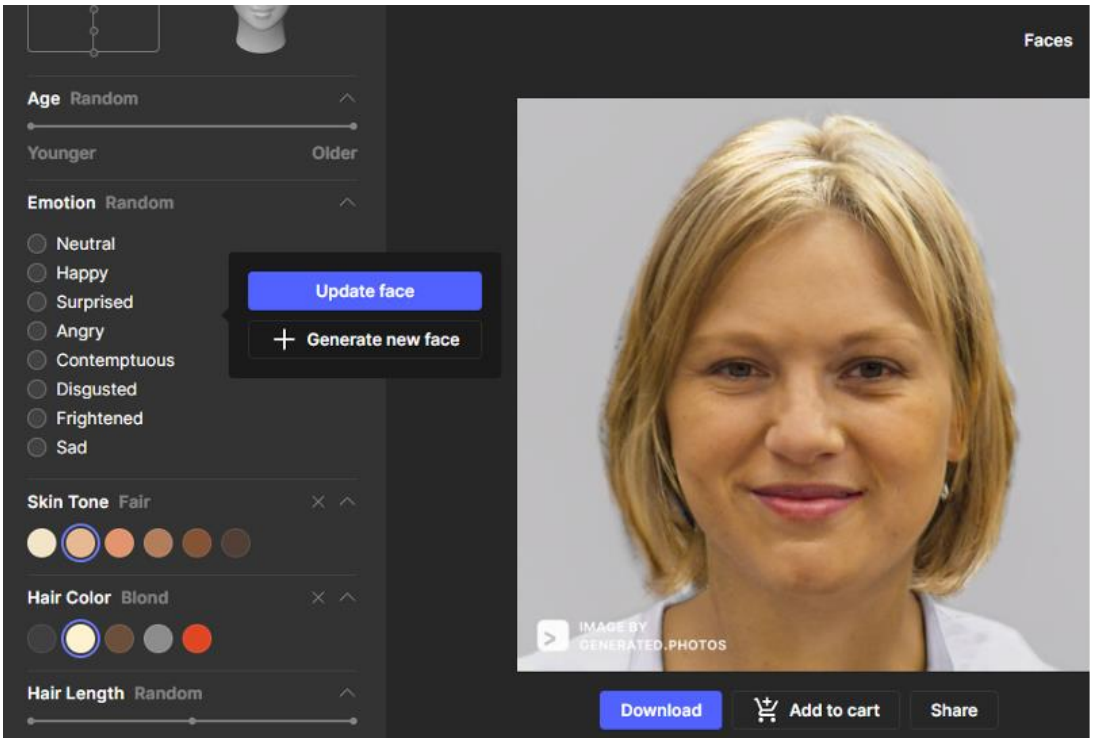

Fig. 1 Générateur de visages de Generated.photos (https://generated.photos/face-generator/new)

Mais ces services ne sont que la partie triviale de l'usage des GAN. En effet, ce ne sont plus seulement de simples visages clonés que la société japonaise *Datagrid* propose depuis 2019, mais des représentations de corps humains synthétiques complets et "animables" qui doivent notamment permettre de remplacer les photos de mannequins dans les magazines de modes ou les défilés virtuels.

Fig. 2 Vidéo de démonstration de la société Datagrid (https://www.youtube.com/watch?v=8siezzLXbNo)

Autre exemple, la société Adobe, très en pointe sur le traitement de l'image via l'IA grâce à sa suite Adobe Sensei, a présenté lors de sa conférence annuelle d'octobre

2021 plusieurs technologies utilisant les GAN pour, par exemple, appliquer la pose spécifique d'un sujet à un autre sujet ou encore modifier l'expression d'une personne sur un cliché.[7]

Fig. 3 Vidéo de démonstration du Project Morpheus d'Adobe (https://www.youtube.com/watch?v=bpABm6XcRMw)

Plus révélatrice encore est l'approche de la société anglaise *Snap*, la maison mère de Snapchat, qui a racheté ces deux dernières années une vingtaine de startups dans les secteurs de l'IA et du Big Data avec pour objectif d'élaborer et diffuser des applications de réalité virtuelle ou augmentée transparentes pour l'utilisateur[8]. Ainsi *Ariel AI* développe une technologie pour smartphone qui permet de "superposer" un corps à un autre en temps réel (*Real-Time 3D reconstruction*) ou d'en générer un à partir de modèles anthropomorphes (*data-driven human modeling*) puis de l'insérer directement dans une scène "*live*", ouvrant ainsi la voie a des hypertrucages diffusés en direct.

Fig. 4 Animation d'un avatar de corps complet (puppeteering) par la startup Ariel.ai (https://www.youtube.com/watch?v=aQ4shIsQabo)

S'il s'agit ici de technologies de modélisation 3D plutôt que de GAN, ces derniers sont à envisager comme la cerise sur le gâteau puisqu'ils permettront l'incrustation de visages hyper réalistes sur les corps virtuels générés. Le rachat de Voca.ai, une startup spécialisée dans les voix synthétiques (cf. ci-dessous), complète d'ailleurs l'ensemble et vient valider l'idée d'une offensive globale de Snapchat sur les hypertrucages et la "fabrication" d'avatars les plus proches possibles de l'humain.

Facebook suit évidemment la même voie, comme l'indiquait récemment Mark Zuckerberg en évoquant le prochain lancement d'un metavers, "*un Internet incarné, où au lieu de simplement regarder le contenu, vous êtes dedans"*. Et de préciser :"*Je pense qu'au cours des cinq prochaines années, dans le prochain chapitre de notre entreprise, nous ferons la transition entre le fait que les gens nous voient principalement comme une entreprise de médias sociaux et le fait que nous soyons une entreprise de metaverse".*[9] Cette direction est désormais confirmée avec le changement de nom de Facebook en Meta[10].

A l'instar de ce qu'annonçait l'auteur de science-fiction Neil Stephenson en forgeant le terme « *metaverse* » pour son roman "*Le samouraï virtuel*", il s'agit donc bien de créer un double virtuel du monde réel dans lequel chacun d'entre nous disposera d'un ou plusieurs avatars plus ou moins réalistes. Il permettra de créer les meilleurs conditions d'interactions interpersonnelles virtuelles possibles (salles de réunions, magasins, visites culturelles, rencontres …) et aura "incidemment" pour effet de démultiplier les capacités de ventes d'espaces de ces méga-régies publicitaires que sont les réseaux sociaux, ainsi que celles des biens en ligne (œuvres, objets, biens immobiliers) grâce au *développement concomitant des NFT*[11]. Car, à l'instar des jeux multi-joueurs en ligne il y aura évidemment des metavers multiples et concurrent et cela est en lien direct avec les technologies du faux puisque c'est au sein des équipes de recherche

de ces firmes qu'elles s'élaborent. Bien entendu, ces univers parallèles seront eux mêmes soumis aux tensions et interactions entre vrai, faux et *fake*. En effet, même si un avatar est un double virtuel, donc nécessairement un faux par rapport à l'individu réel qu'il incarne, il pourra avoir statut d'avatar "officiel", à savoir validé par l'individu qu'il représente et la plateforme qui l'accueille (les technologies de *blockchain* auront probablement un rôle important à jouer ici aussi). Le même individu pourra et devra alors très probablement créer des avatars *fake*, des *sockpuppets*[12], qui lui permettront de retrouver un minimum d'anonymat, tout comme l'on dispose souvent de plusieurs profils sur les réseaux sociaux. Il va sans dire que pour d'autres individus (parfois les mêmes) ces avatars seront créés dans le seul but de tromper les quidams rencontrés dans le metavers.

Point important qui n'avait pas encore été souligné au sujet des *deepfakes*, il est possible depuis 2018 de mettre en œuvre une substitution de visage (*face swapping*) sur une diffusion en temps réel avec un résultat souvent bluffant. Le logiciel le plus utilisé pour cela, *DeepFaceLive*, est gratuit et s'installe sous Windows 10. Il ne nécessite même plus les longues phases d'entraînement d'algorithmes sur sets d'images (qui peuvent prendre entre 3 et 10 jours) puisqu'il est possible d'utiliser des modèles de visages "prêts à l'emploi", déjà compilés par d'autres utilisateurs.

Les entreprises des *synthetic media* spécialisées en ce domaine ont d'ores et déjà mis au point des modèles économiques spécifiques à l'instar d'*Hour One*, qui rémunère des personnes pour qu'elles cèdent les droits sur leur visage. Un catalogue de ces visages est ensuite proposé aux clients qui peuvent l'exploiter pour certains usages déjà répertoriés (et d'autres à imaginer). Cela va de l'accompagnement de visites de biens immobiliers en ligne à la présentation vidéo de rapports financiers, en passant par les cours de langue.

Des voix

Si les *deepfakes* font la part belle aux images, le domaine de l'audio innove au moins autant autour de deux axes, celui des sons et de la musique synthétique et celui des voix. Ce dernier nous intéresse particulièrement puisqu'il permet de donner la parole aux avatars et la société *Hour.one* en est encore un bon exemple. Cette startup israëlo-américaine propose, en parallèle de son catalogue de visages, un catalogue de voix alimenté sur le même principe de cession des droits par des individus contre rémunération. Une fois clonées, ces voix sont couplées aux visages choisis par le client et la solution d'Hour One peut alors générer une quantité infinie de séquences d'une personne récitant n'importe quel texte, dans n'importe quelle langue. La société Berlitz, un client d'Hour One spécialisé dans l'enseignement des langues, indique générer ainsi des centaines de vidéos en quelques minutes. "*Nous remplaçons le studio (…). Un être humain n'a pas besoin de perdre son temps à filmer*".

Autre startup en vue dans ce secteur, *Synthesia* rémunère des acteurs pour enregistrer leur voix et propose également de nombreux avatars multilingues. Son interface est très orientée vers l'utilisateur final qui peut choisir parmi une bibliothèque d'acteurs (ou téléverser ses propres modèles) puis créer une scène en ajoutant des composants tels que des meubles, des objets (bien souvent *générés également par des GAN*), du texte, des images et créer ainsi une vidéo sans compétences spécifiques. Là encore les langues parlées sont multiples comme le montre ce clip

dans lequel, grâce à la technologie de Synthesia, David Beckham évoque la lutte contre la malaria en neuf langues.

Fig. 5 David Beckham évoque la lutte contre la malaria en neuf langues avec la technologie de Synthesia.ai (https://www.youtube.com/watch?v=QiiSAvKJIHo)

Chez *Sonantic* ou *CyberVoice*, on poursuit le même objectif appliqué aux personnages de jeux vidéos. Là aussi on clone les voix des interprètes afin de pouvoir les réutiliser à loisir dans les scènes de jeu. Ce qui ne va pas sans inquiéter les acteurs, souvent semi-professionnels, quant à leur "utilité" une fois leur voix clonée. Comme l'explique l'une d'elle "*imaginez que vous devenez un personnage aimé par beaucoup mais que vous n'avez pas fait une seule chose pour contribuer à ce rôle. Zéro créativité de la part de l'acteur. Zéro épanouissement. Zéro art*."[13] Leur inquiétude porte également sur l'utilisation qui sera faite de leur voix. Que se passe t-il si l'éditeur s'en sert pour véhiculer des idées ou des mots que son possesseur n'approuve pas ? Les questions liées à l'éthique et aux droits associés à la voix d'un individu et plus globalement à toute utilisation d'un clonage de ce qui le constitue vont entraîner dans les années à venir d'inévitables batailles judiciaires. Ainsi, l'utilisation de la voix synthétique d'Anthony Bourdain, chef cuisinier star décédé en 2018, pour lui faire prononcer des paroles "inédites" dans un film lui rendant hommage, a déjà créé la polémique, tant du côté des spectateurs qui se sont sentis trompés que des ayants-droits qui n'avaient a priori pas donné leur accord[14].

L'usage de ces technologies va changer d'échelle dans les mois à venir puisque, pour la première fois, un film sera entièrement doublé avec des voix synthétiques créées à partir de celles des acteurs anglo-saxons. Ainsi, lorsque le *thriller* américain *Every Time I Die* (sorti en 2019) sera diffusé dans les salles en Amérique du Sud, les spectateurs entendront les interprètes originaux parler en espagnol et en portugais grâce à la technologie de la société *Deepdub*.[15]

Si plusieurs startups de ce domaine mettent en avant sur leur site web une déclaration éthique indiquant les conditions d'utilisation des voix qu'elles clonent, il est clair qu'elles ne le font pas toutes. Par ailleurs, les programmes de "*voice cloning*" sont très nombreux et en accès libre sur *Github*… Autant dire que les prochaines années risquent d'être un *eldorado* pour les startups peu regardantes et leur clients, et un champ de bataille judiciaire sans fin d'où émergeront des lois et jurisprudences à la fois nécessaires et difficiles à faire respecter.

Des corps synthétiques pour alimenter l'IA

Si, comme on l'a vu, il faut des centaines d'images d'une personne cible pour obtenir une très bonne correspondance, les choses pourraient cependant évoluer rapidement. En effet, les possibilités atteintes par les *deepfakes*, déjà impressionnantes, sont encore plus avancées que ne le laissent entrevoir les exemples déjà présentés. Ainsi, a l'instar d'Ariel.ai évoquée précédemment, la société *Datagen*, propose, des représentations de corps humains synthétiques complets à animer. Pour les générer, la société scanne de vraies personnes payées pour cela et utilise ensuite ces données brutes pour les faire passer par une série d'algorithmes, qui créent des représentations en 3D de leur corps, visage, yeux et mains. Son objectif est toutefois différent de celui des entreprises précédentes puisqu'il s'agit ici de proposer des sets de données d'individus virtuels totalement configurables et combinables en terme de variance (âge, sexe, taille, poids). Ce "produit" doit permettre à leurs clients de constituer des panels d'avatars et d'étudier leur comportement en les intégrant à des simulations où ils joueront le rôle d'humains. Il s'agira par exemple de suivre et comprendre leurs mouvements de corps lors d'un passage en magasin sans caisse, leurs expressions faciales afin de monitorer la vigilance des conducteurs de voitures intelligentes, ou encore l'usage qu'ils feront d'une manette de jeu.

Fig. 6 Simulation permettant de travailler sur l'ergonomie d'une manette de jeu - Datagen (https://www.youtube.com/watch?v=6JkUk3C9bHI)

Même idée pour *Synthesis.ai* qui met à disposition des sets de visages avec une variabilité incluant par exemple le port de lunettes, de masque, les expressions, l'éclairage,…

Fig. 7 Exemple de "sets de visages" proposés par Synthesis.ai (https://www.youtube.com/watch?v=NoGzuLcaDOY)

https://youtu.be/NoGzuLcaDOY

Le dispositif, baptisé *HumanAPI*, se présente sous la forme d'une interface applicative utilisable par d'autres entreprises qui peuvent ainsi l'intégrer à leurs propres services et logiciels. La donnée synthétique anthropomorphe « *as-a-service* » en somme. Et le PDG et fondateur de Synthesis.ai, Yashar Behzadi, de préciser : "*HumanAPI permet également toutes sortes de nouvelles opportunités pour nos clients, notamment des assistants intelligents d'IA, des coachs de fitness virtuels et, bien sûr, le monde des applications du metavers.*"[16]

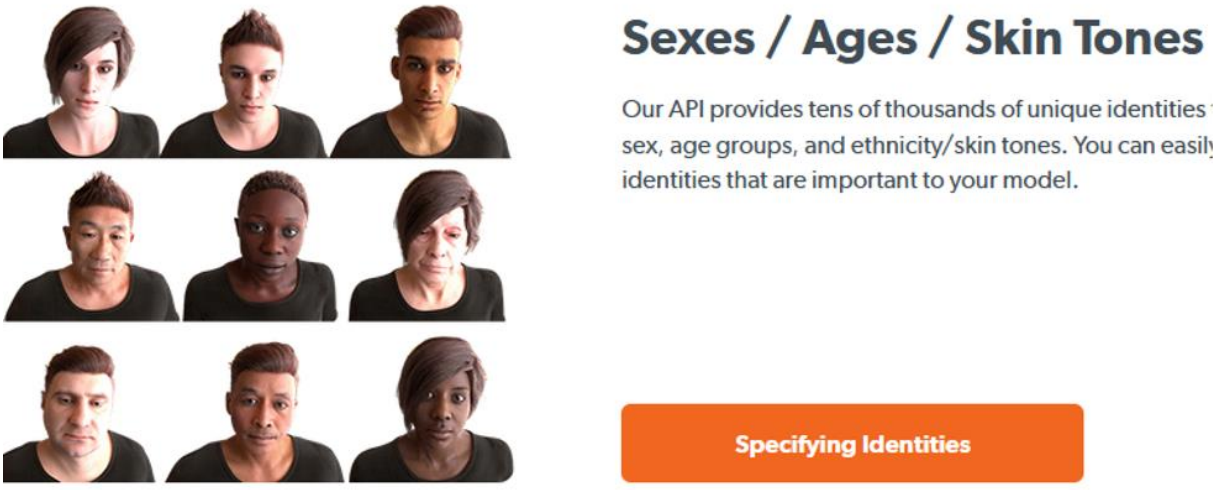

Fig. 8 Page de présentation de l'interface HumanAPI (https://synthesis.ai/api/)

Ce que l'on voit émerger ici est clairement un nouveau marché, celui des données synthétiques dont les usages sont innombrables[17] et qui vont permettre d'alimenter les algorithmes d'apprentissage profond, notamment les GAN, en images hyperréalistes à bas coût plutôt que de les collecter difficilement dans le monde réel du fait des réglementations sur la confidentialité et le respect de la vie privée.

En effet, les données synthétiques sont "vierges" et peuvent être utilisées pour créer des ensembles de données plus diversifiés. On peut, comme *Synthesis.ai*, générer facilement des visages bien étiquetés par âges, formes, IMC, couleur de peau, afin de créer un système de reconnaissance faciale qui fonctionnera pour toutes les populations. Ce qui ne veut pas dire que les algorithmes utilisés pour générer ces faux n'ont aucun biais. Comment s'assurer par exemple que l'expression que l'on a modélisée initialement sur un humain "véritable", comme étant de la joie ou de la colère était bien ce qu'il voulait exprimer et surtout qu'elle sera interprétée comme telle par ceux qui la verront ? La complexité des expressions humaines est telle que la réduire à un catalogue semble pour le moins problématique et risque de générer des choix et des actions basés sur des interprétations erronées du réel.

Notons que le marché des données synthétiques est en pleine expansion et ne touche pas seulement la création d'avatars mais plus globalement tout secteur d'activité ayant besoin de *datasets* complexes à collecter ou onéreux, afin d'être utilisées pour simuler des données dont il ne dispose pas.

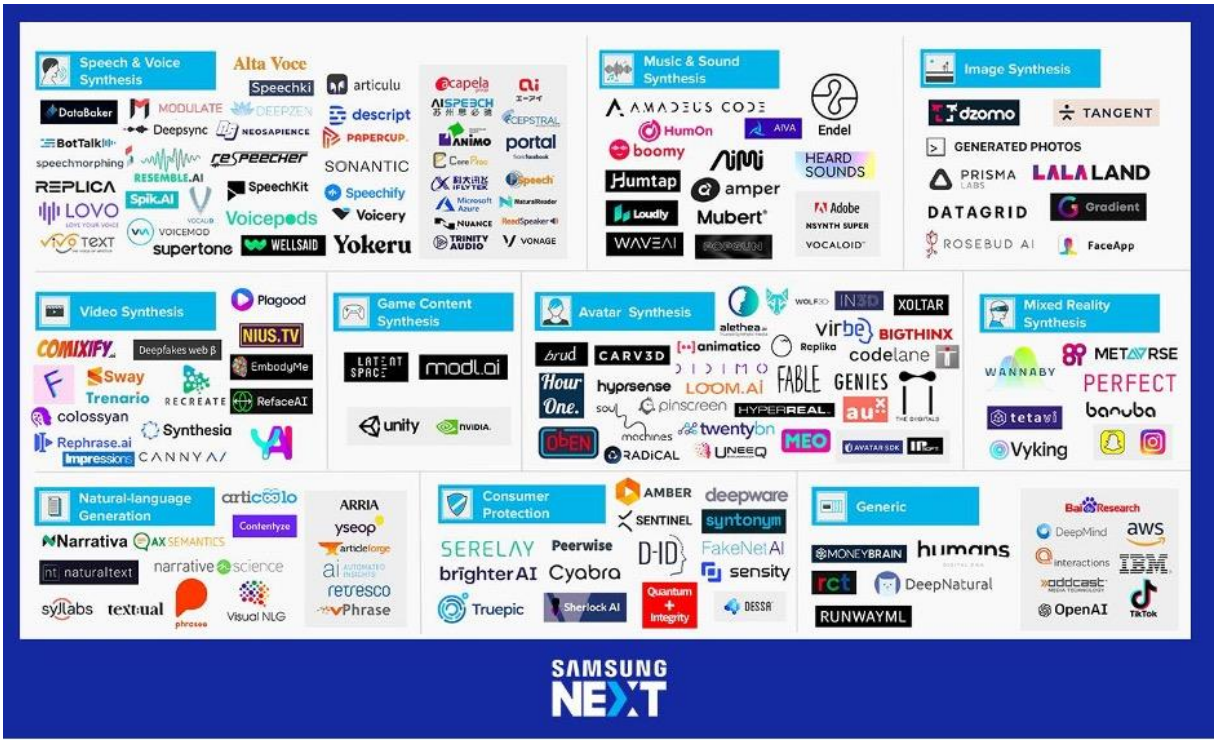

Fig. 9 Cartes des entreprises du "synthetic medias" (pas Samsung Next Ventures)

Il peut s'agir comme on l'a vu d'une population de clients plus diversifiée mais aussi de transactions impliquant des problématiques de respect de la vie privée. Les données synthétiques sont également très utilisées pour gérer des questions de confidentialité. Ainsi la société *Mostly.ai* travaille avec des sociétés financières, de télécommunications et d'assurance pour fournir des jeux de données clients répartis différemment des vraies mais sur un même volume, permettant ainsi aux entreprises

de partager leur base de données clients avec des fournisseurs extérieurs tout en restant en conformité avec la loi.

Fig. 10 Présentation des sets de données synthétiques de la société Mostly.ai (https://www.youtube.com/watch?v=NjpkeiUcc5c)

Quoiqu'il en soit, ce nouveau marché se construit sur une capacité croissante à entraîner des algorithmes de *deep learning* à partir de jeux de données virtuels, c'est à dire à créer du faux vraisemblable avec du faux crédible (ou inversement). Si cela s'avère être une solution intéressante lorsqu'il s'agit de modéliser et anticiper le comportement d'un système industriel ou d'anonymiser un portefeuille clients, il nous semble qu'il n'en va pas de même avec la "matière anthropomorphe". Le risque étant de tenter, par ces modélisations, de prévoir l'imprévisible à savoir le comportement humain.

Toute image est bonne à prendre

Les représentations du visages ou des parties du corps humain ne sont cependant pas les seuls à pouvoir être exploitées par ces algorithmes pour lesquels une image en vaut une autre et qui sont donc capables de tirer parti de n'importe quel jeu de données visuel, quoi qu'il représente. Ainsi les expérimentations se multiplient dans de nombreuses directions. En mai dernier par exemple, une étude[18] menée par le chercheur Bo Zhao et son équipe, à l'université de Washington, montrait qu'il était possible de créer grâce aux GAN de vraies-fausses images satellites de villes.

Appliqué au domaine de la cartographie, l'algorithme apprend essentiellement les caractéristiques des images satellite d'une zone urbaine, puis génère une fausse image en introduisant les caractéristiques de l'image satellite apprise dans une carte de référence différente, de la même manière que les filtres d'image populaires peuvent reproduire les caractéristiques d'un visage sur un autre.

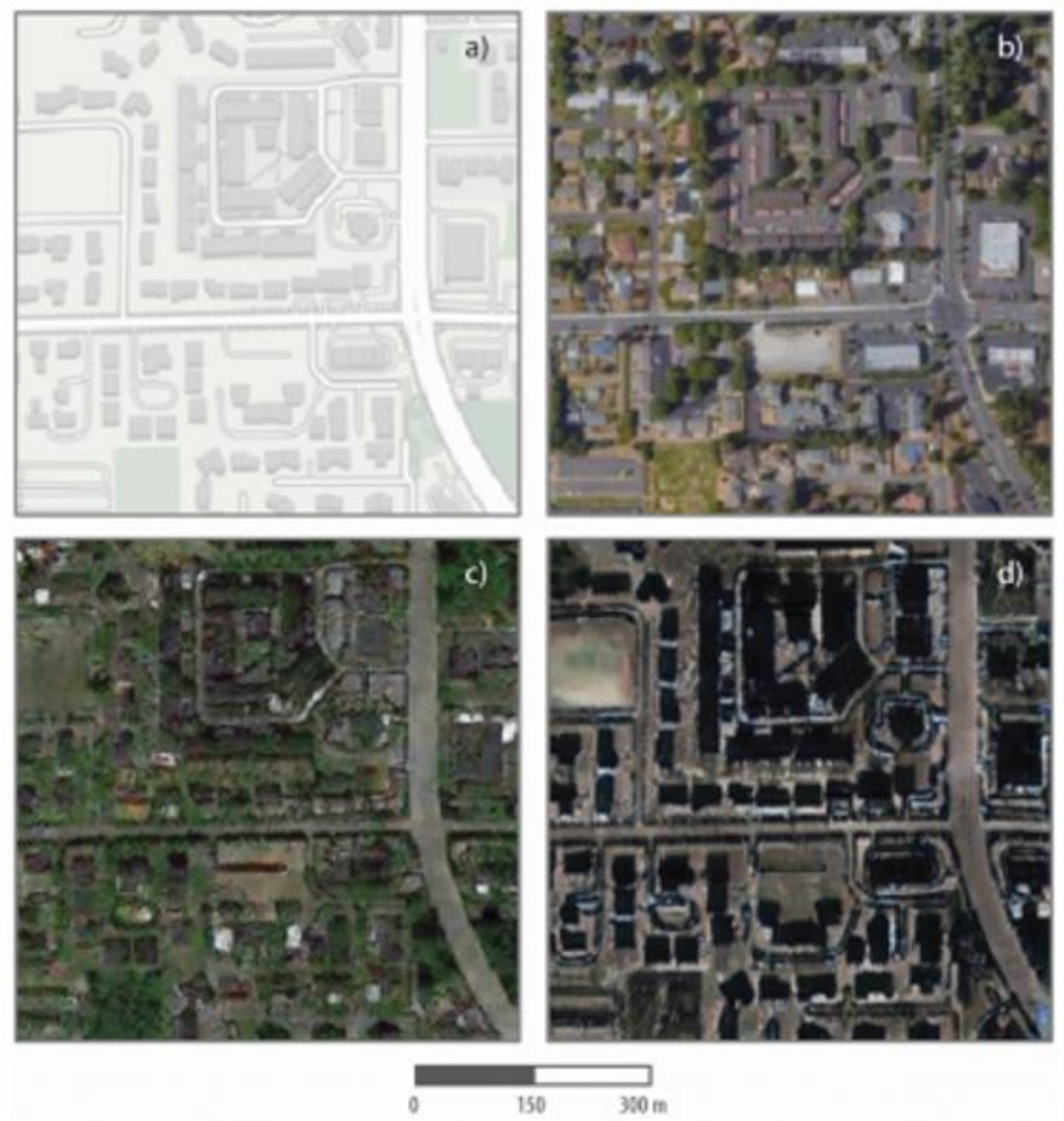

Fig. 11 Les fausses images satellitaires c) et d) ont été générées à partir des images a) et b)

Ici, les chercheurs ont donc combiné les images de trois villes, Tacoma, Seattle et Pékin en créant de nouvelles images d'une ville à partir des caractéristiques des deux autres. Ils ont désigné Tacoma comme ville de référence et y ont ensuite intégré les caractéristiques géographiques et urbaines de Seattle et de Pékin pour créer un *deepfake* de celle-ci. Les possibilités offertes ici sont vertigineuses et Bo Zhao, dont le but était autant d'explorer les possibilités de création que de détection de fausses images satellites de conclure : "*cette étude vise à encourager une compréhension plus holistique des données et des informations géographiques, afin que nous puissions démystifier la question de la fiabilité absolue des images satellites ou d'autres données géospatiales (…) Nous voulons également développer une réflexion plus orientée vers l'avenir afin de prendre des contre-mesures telles que la vérification des faits lorsque cela est nécessaire*"[19].

L'imagerie satellite étant depuis longtemps la base des prévisions météorologiques, le champ d'action s'est vite étendu en ce sens et *plusieurs équipes de chercheurs* utilisent déjà les GAN dans le but de les améliorer en enrichissant ainsi les

modélisations. L'un de ces projets, mené en collaboration par la société anglaise *Deepmind* et le *Met Office* (le Météo France anglais), a vu, lors d'une comparaison en aveugle avec les outils existants, plusieurs dizaines d'experts juger que les prévisions données par le modèle GAN étaient meilleures pour l'emplacement, l'étendue, le mouvement et l'intensité de la pluie, et ce dans 89 % des cas[20].

Mais l'imagerie satellite n'est pas la seule à attirer les expérimentations des chercheurs, l'imagerie médicale voit elle aussi se multiplier les études visant à évaluer l'utilisation d'images synthétiques de qualité, peu coûteuses et non invasives (*deepfakes* d'images radiographiques, de scanner, d'IRM…) pour entrainer d'autres systèmes de détection utilisant également le *deep learning*, a l'instar des corps synthétiques déjà évoqués. Les GAN les plus performants sont capables de générer des images médicales réalistes qui peuvent tromper des experts entraînés, même si, comme le soulignent les auteurs d'une de ces études, "*aucun GAN n'est capable de reproduire toute la richesse d'un jeu de données médicales*"[21]. Pour l'instant…

Du texte et du code

Si l'image, photographie ou vidéo, présente l'usage le plus spectaculaire de cette nouvelle industrie du faux, le texte n'est pas loin derrière. Les possibilités de traitement du langage naturel (*natural language processing* ou NLP en anglais) se sont en effet démultipliées avec l'arrivée de nouveaux modèles avancés intégrant l'intelligence artificielle et plus spécifiquement les réseaux de neurones. On y trouve *BERT*, proposé par Google, mais le plus prometteur est GPT-3 (bientôt GPT-4) d'*OpenAI*, une entreprise fondée notamment par Elon Musk. GPT-3 (pour *Generative Pre-Trained Transformer* 3) est un modèle alimenté par 175 milliards de paramètres, c'est-à-dire de valeurs qu'un réseau de neurones essaye d'optimiser pendant l'entraînement (son prédécesseur GPT-2 n'en n'avait "que" 1,5 milliards). Le modèle est donc conçu pour générer du texte en utilisant des algorithmes déjà entraînés, c'est à dire nourris d'un corpus de références collectées sur le web (l'intégralité de la Wikipedia par exemple). Cet entraînement lui permet, via une analyse sémantique, de "comprendre" la mécanique d'une langue (essentiellement l'anglais pour l'instant). Une fois passée cette étape, lorsqu'on fournit à l'algorithme un extrait de texte, par exemple une phrase d'introduction, celui-ci va tenter de le compléter en prédisant les mots qui pourraient faire sens pour l'utilisateur, comme on peut le constater en testant le service *Talktotransformer* (sous GPT-2).

Les possibilités sont alors innombrables et limitées par la seule imagination :

- *rédiger un guide* pour des réunions plus efficaces,
- *créer des chatbots* pour répondre à des utilisateurs ou des clients,
- *écrire des articles* pour alimenter un blog sur la productivité (1 personne sur 26000 s'apercevra de la supercherie volontaire),
- *écrire des textes* "à la manière de" (ici Jerome K. Jerome),
- demander à GPT-3 de *répondre à une question d'ordre médicale*,
- *générer une aventure* de jeu de rôle en ligne en temps réel (ici du *jeu de rôle textuel*),
- *créer des interfaces utilisateurs* à partir d'une simple description textuelle.

Mais ces exemples ne donnent qu'une idée limitée des usages potentiels de ce type d'algorithmes et l'on trouve déjà sur le site *gpt3demo* plus de deux-cent exemples de mises en œuvre dans des domaines aussi variés que la *pensée créative*, le *recrutement*, la *poésie* ou… la *création de phrases d'accroche pour Tinder*. Notons que la génération d'image évoquée précédemment n'est jamais très loin puisque *l'outil DALL.E*, présent dans cet annuaire et développé également par OpenAI, permet de créer des images à partir d'un texte ou de la voix.

Fig. 12 Exemples de créations par DALL.E en réponse à la question "une chaise imitant un avocat"

Parmi toutes ces possibilités déjà existantes il en est une qui est à considérer comme la pierre angulaire de l'essor imminent des technologies du faux. C'est la capacité de GPT-3 à générer du code informatique. Présentée en avril dernier, cette fonction donne des résultats impressionnants et Microsoft l'a déjà intégrée à ses offres d'entreprise Power Apps[22] et plus récemment Azure[23] afin de permettre à chacun de développer en *no code/low-code*. On peut se faire une idée de la puissance de cette fonctionnalité en visionnant la vidéo de présentation proposée par OpenAI, où l'on voit deux développeurs créer une campagne d'emails et l'envoyer via le service Mailchimp, formater un texte dans Word ou coder un jeu uniquement à la voix.

Fig. 13 Présentation de l'outil de génération de code via IA d'OpenAI (https://www.youtube.com/watch?v=SGUCcjHTmGY)

L'impression laissée par cette vidéo est celle d'une magie en acte, d'une parole qui devient créatrice par l'intermédiaire d'une IA invisible ou qui le deviendra bientôt. Ou quand le logiciel laisse la place au Logos…
Ce qu'il faut donc absolument saisir c'est que dans un futur proche GPT-3 et ses concurrents à venir (notamment *MT-NLG* de Microsoft et NVIDIA) vont mettre cette parole créatrice à la disposition de tous. C'est la promesse des technologies *No code* ou *low-code* dont l'utilisation sera responsable de 65% de l'activité de développement en 2024 d'après le Magic Quadrant de Gartner[24] et dont le marché mondial devrait générer un revenu de 187 milliards de dollars d'ici 2030, contre 10 milliards de dollars en 2019[25].

La commoditisation est en marche et le territoire à explorer infini…

Conclusion

Si ce tour d'horizon des technologies du faux nous laisse percevoir l'ampleur des changements à venir dans nos usages quotidiens, il n'avait pour objectif que de tenter de circonscrire le sujet et de nombreuses dimensions n'ont donc pas été abordées. Tout d'abord la véracité et la pertinence des faux générés. En effet, les *deepfakes* sont (heureusement) loin d'être toujours crédibles et leur potentiel en tant qu'artifice en est dès lors compromis. Il peuvent également s'avérer nuisibles, comme lorsqu'un *chatbot* médical utilisant GPT-3 conseille à un (faux) patient de se suicider[26]. Par ailleurs la question des biais qui y sont introduits, volontairement ou non, par les développeurs et les jeux de données utilisés reste sensible et complexe à traiter : sur quels critères choisir les biais à limiter ou à privilégier ? Qui est légitime pour en juger ? Faut-il un vote démocratique pour en décider ?

Autre dimension essentielle non traitée ici, les innombrables risques que ces technologies font courir à notre société en ce que le faux peut être élaboré dans l'intention de nuire, devenant alors du *fake* créé pour toutes les raisons, bonnes ou mauvaise, amenant un acteur ou une entité à vouloir en manipuler d'autres. Nous n'avons pas encore développé ces aspects négatifs mais ils peuvent déjà être aisément déduits des exemples que nous avons donnés.

Une troisième dimension est celle des impacts que les technologies du faux et leur usage en mode *fake* aura sur les organisations publiques ou privées et, partant, des techniques de détection et des mesures (et contre-mesures) à déployer pour en enrayer la propagation. C'est ce sujet et plus précisément ce qu'il implique dans les sphères de l'intelligence économique et de l'analyse du renseignement, que nous aborderons dans un prochain article.

Christophe Deschamps est consultant et formateur indépendant sur les thématiques de veille stratégique, d'intelligence économique et de gestion des connaissances. Il est l'auteur du blog www.outilsfroids.net, consacré à ces mêmes thèmes. Il est par ailleurs doctorant au CEREGE.

Bibliographie

[1] François-Bernard Huyghe (2016) Désinformation : armes du faux, lutte et chaos dans la société de l'information. In : Sécurité globale, vol. 6, n° 2, p. 63–72. En ligne : https://www.cairn.info/revue-securite-globale-2016-2-page-63.htm.
[2] LeCun, Yann; Bengio, Yoshua; Hinton, Geoffrey (2015) Deep learning. In : Nature, vol. 521, n° 7553, p. 436–444. DOI: 10.1038/nature14539
[3] Goodfellow, Ian J.; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil et al. (2014) Generative Adversarial Networks. En ligne : https://arxiv.org/pdf/1406.2661.
[4] Holubowicz, Gérard (2021) L'histoire des deepfakes. En ligne : https://journalism.design/chapitre-1-histoire-des-deepfakes/, consulté le 10 novembre 2021.
[5] Cole, Samantha (2017) AI-Assisted Fake Porn Is Here and We're All Fucked. In : VICE, 12 novembre 2017. En ligne : https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn, consulté le 11 novembre 2021.
[6] Thies, Justus; Zollhofer, Michael; Stamminger, Marc; Theobalt, Christian; Niessner, Matthias (2016) Face2Face: Real-Time Face Capture and Reenactment of RGB Videos: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR): IEEE. En ligne : http://openaccess.thecvf.com/content_cvpr_2016/papers/Thies_Face2Face_Real-Time_Face_CVPR_2016_paper.pdf.
[7] Ahmed, Arooj (2021) Adobe Revealed Several New AI Powered Features At The Annual MAX Conference, Here Are Three Best Of Them / Digital Information World. En ligne : https://www.digitalinformationworld.com/2021/10/adobe-revealed-several-new-ai-powered.html, consulté le 11 novembre 2021.
[8] Pimenta, Joana (2021) Snapchat rachète une entreprise par mois. In : Siècle Digital, 4 juin 2021. En ligne : https://siecledigital.fr/2021/06/04/acquisitions-snap-snapchat/, consulté le 11 novembre 2021.
[9] Reisacher, Appoline (9/30/2021) Metaverse : tout savoir sur cet univers virtuel qui attire les géants de la tech. In : BDM, 9/30/2021. En ligne : https://www.blogdumoderateur.com/metaverse-univers-virtuel-attire-geants-tech/, consulté le 11 novembre 2021.
[10] Neveu, Louis (2021) Pourquoi Facebook devient Meta ? En ligne : https://www.futura-sciences.com/tech/actualites/facebook-facebook-devient-meta-94538/, consulté le 11 novembre 2021.
[11] Les *Non Fongible Tokens* (NFT) ou Jetons non fongibles sont des certificats d'authenticité attribués aux créations numériques. Ils permettent de s'approprier une œuvre et d'être authentifié comme propriétaire unique via la technologie de *blockchain*.
[12] Le terme de *sockpuppet* (en français « marionnette ») est utilisé par les hackers, pentesters et investigateurs OSINT pour désigner les faux profils créés pour mener leurs activités anonymement.
[13] Hart, Aimee (6/26/2021) Voice AI is scary good now. Video game actors hate it. In : Input, 6/26/2021. En ligne : https://www.inputmag.com/gaming/video-game-voice-ai-human-actors-witcher-3-mod-controversy, consulté le 11 novembre 2021.
[14] O'Brien, Matt; Ortutay, Barbara (7/17/2021) Why the Anthony Bourdain voice cloning creeps people out. In : Associated Press, 7/17/2021. En

ligne : https://apnews.com/article/anthony-bourdain-documentary-voice-cloning-technology-1dae37f748a22c946e2193fbb00ccc11, consulté le 11 novembre 2021.
[15] Gamerman, Ellen (2021) The Rise of the Robo-Voices. In : The Wall Street Journal, 10 juillet 2021. En ligne : https://www.wsj.com/articles/the-rise-of-the-robo-voices-11633615201, consulté le 11 novembre 2021.

[16] AI, Synthesis (2021) Synthesis AI Launches HumanAPI to Create Millions of Photorealistic Digital Humans, On-Demand, 11 septembre 2021. En ligne : https://www.prnewswire.com/news-releases/synthesis-ai-launches-humanapi-to-create-millions-of-photorealistic-digital-humans-on-demand-301419311.html, consulté le 15 novembre 2021.
[17] Cf. cette étude par Samsung Next Ventures, Synthetic Media Landscape 2020. En ligne : https://www.syntheticmedialandscape.com/, consulté le 11 novembre 2021.

[18] Zhao, Bo; Zhang, Shaozeng; Xu, Chunxue; Sun, Yifan; Deng, Chengbin (2021) Deep fake geography? When geospatial data encounter Artificial Intelligence. In : Cartography and Geographic Information Science, vol. 48, n° 4, p. 338–352. DOI: 10.1080/15230406.2021.1910075
[19] Eckart, Kim (2021) A growing problem of 'deepfake geography': How AI falsifies satellite images. UW News, éd. En ligne : https://www.washington.edu/news/2021/04/21/a-growing-problem-of-deepfake-geography-how-ai-falsifies-satellite-images/, consulté le 11 novembre 2021.
[20] Ravuri, Suman; Lenc, Karel; Willson, Matthew; Kangin, Dmitry; Lam, Remi; Mirowski, Piotr et al. (2021) Skilful precipitation nowcasting using deep generative models of radar. In : Nature, vol. 597, n° 7878, p. 672–677. DOI: 10.1038/s41586-021-03854-z
[21] Skandarani, Youssef; Jodoin, Pierre-Marc; Lalande, Alain (2021) GANs for Medical Image Synthesis: An Empirical Study. En ligne : https://arxiv.org/pdf/2105.05318.
[22] Lardinois, Frederic (5/25/2021) Microsoft uses GPT-3 to let you code in natural language. In : TechCrunch, 5/25/2021. En ligne : https://techcrunch.com/2021/05/25/microsoft-uses-gpt-3-to-let-you-code-in-natural-language/, consulté le 11 novembre 2021.
[23] Aballéa, Arthur (2021) Microsoft facilite l'accès à GPT-3 sur Azure via l'API OpenAI. In : BDM, 11 février 2021. En ligne : https://www.blogdumoderateur.com/microsoft-facilite-acces-gpt-3-azure-via-api-openai/, consulté le 11 novembre 2021.
[24] Low-Code Is the Future – OutSystems Named a Leader in the 2019 Gartner Magic Quadrant for Enterprise Low-Code Application (2019). In : Bloomberg, 8 décembre 2019. En ligne : https://www.bloomberg.com/press-releases/2019-08-12/low-code-is-the-future-outsystems-named-a-leader-in-the-2019-gartner-magic-quadrant-for-enterprise-low-code-application, consulté le 11 novembre 2021.
[25] Prescient & Strategic Intelligence Private Limited, éd. (2020) Low-Code Development Platform Market Research Report: By Offering, Deployment Type, Enterprise, Vertical – Global Industry Analysis and Growth Forecast to 2030. En ligne : https://www.researchandmarkets.com/reports/5184624/low-code-development-platform-market-research, consulté le 11 novembre 2021.
[26] Daws, Ryan (2020) Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves. En ligne : https://artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/, consulté le 11 novembre 2021.